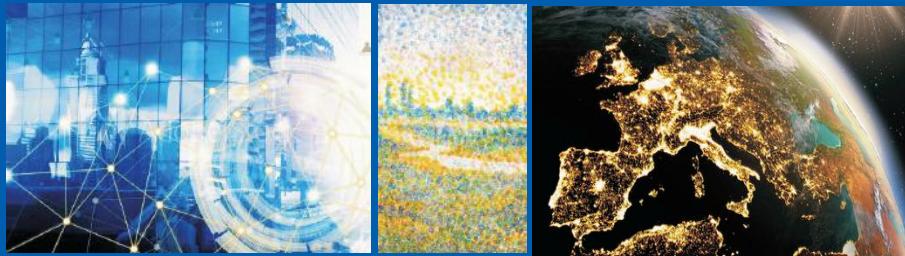


# Insee Méthodes



N° 131  
October 2018

## Handbook of Spatial Analysis

Theory and Application with R

**Insee - Eurostat**

Directed by Vincent LOONIS - Coordinated by Marie-Pierre de BELLEFON



# **HANDBOOK OF SPATIAL ANALYSIS**

**Theory and practical application with R**

**This project has received funding from the European Statistical Programme 2013-2017 under  
the ESS Action Integration of statistical and geospatial information with grant agreement No  
08143.2015.001-2015.714**

Direction	Vincent Loonis
Coordination	Marie-Pierre de Bellefon
Editorial consistency	Vianney Costemalle, Maëlle Fontaine
Contribution	<p><i>INSEE:</i>            Pascal Ardilly, Sophie Audric, Marie-Pierre de Bellefon,            Maël-Luc Buron, Eric Durieux, Pascal Eusébio, Cyril            Favre-Martinoz, Jean-Michel Floch, Maëlle Fontaine, Laure            Genebes, Ronan Le Gleut, Raphaël Lardeux-Schutz, David            Lévy, Vincent Loonis, Ronan Le Saout, Thomas Merly-Alpa,            Auriane Renaud, François Sémécubre</p> <p><i>GAINS (TEPP) and CREST Le Mans Université</i>            Salima Bouayad-Agha</p> <p><i>AgroParisTech, UMR EcoFoG</i>            Eric Marcon</p> <p><i>RITM, Univ. Paris-Sud, Université Paris-Saclay and CREST</i>            Florence Puech</p> <p><i>CESAER, AgroSup Dijon, INRA, Université de Bourgogne</i>  <i>Franche- Comté</i>            Julie Le Gallo, Lionel Védrine</p> <p><i>ENSAI</i>            Paul Bouche, Wencan Zhu</p>
Director of the publication	Jean-Luc Tavernier
Editor	Institut national de la statistique et des études économiques 88, avenue Verdier - CS 700058 92541 Montrouge Cedex <a href="http://www.insee.fr">www.insee.fr</a>
<p>©INSEE Eurostat 2018 "Partial reproduction is authorised provided the source is acknowledged."</p>	

# Table of contents

I

## Part 1: Describing geolocalized data

<b>1</b>	<b>Descriptive Spatial Analysis</b>	<b>3</b>
1.1	Different types of spatial data	4
1.2	Concepts in cartographic semiology	7
1.3	Mapping data with R	10
1.4	Examples of studies using aggregated spatial data	26

<b>2</b>	<b>Codifying the neighbourhood structure</b>	<b>31</b>
2.1	Defining neighbours	32
2.2	Attributing weights to neighbours	42

II

## Part 2: Measuring the importance of spatial effects

<b>3</b>	<b>Spatial autocorrelation indices</b>	<b>51</b>
3.1	What is spatial autocorrelation?	52
3.2	Global measures of spatial autocorrelation	55
3.3	Local measures of spatial autocorrelation	62
3.4	Spatio-temporal indices	68

<b>4</b>	<b>Spatial distribution of points</b>	<b>71</b>
4.1	Framework of analysis: basic concepts	74
4.2	Point processes: a brief presentation	76
4.3	From point processes to observed point distributions	81
4.4	What statistical tools should be used to study spatial distributions?	83
4.5	Recently proposed distance-based measures	95
4.6	Multi-type processes	98
4.7	Process modelling	107

<b>5</b>	<b>Geostatistics</b>	<b>113</b>
5.1	Random functions	114
5.2	Spatial variability	116
5.3	Fitting variogram	122
5.4	Ordinary kriging	127
5.5	Support and change of support	134
5.6	Extensions	136

<b>6</b>	<b>Spatial econometrics - common models .....</b>	<b>149</b>
6.1	What are the benefits of taking spatial, organisational or social proximity into account?	151
6.2	Autocorrelation, heterogeneity and weightings: a review of key points in spatial statistics	152
6.3	Estimating a spatial econometrics model	154
6.4	Econometric limits and challenges	160
6.5	Practical application under R	163
<b>7</b>	<b>Spatial econometrics on panel data .....</b>	<b>179</b>
7.1	Specifications	180
7.2	Estimation methods	186
7.3	Specification tests	189
7.4	Empirical application	190
7.5	Extensions	198
<b>8</b>	<b>Spatial smoothing .....</b>	<b>205</b>
8.1	Kernel smoothing	206
8.2	Geographical smoothing	212
8.3	Implementation with R	217
<b>9</b>	<b>Geographically Weighted Regression .....</b>	<b>231</b>
9.1	Why use geographically weighted regression?	232
9.2	Geographically Weighted Regression	234
9.3	Robust Geographically Weighted Regression	240
9.4	Quality of estimates	246
9.5	A predictive application	247
9.6	Precautions to take	250
<b>10</b>	<b>Spatial sampling .....</b>	<b>255</b>
10.1	General	256
10.2	Constructing primary sampling units of small area and with a constant number of secondary statistical units	257
10.3	How to draw a spatially balanced sample	260
10.4	Comparing methods	268

<b>11</b>	<b>Spatial econometrics on survey data .....</b>	<b>277</b>
11.1	First approach by simulation	280
11.2	Prospects for resolution	288
11.3	Empirical application: the manufacturing sector in Bouches-du-Rhône	291
<b>12</b>	<b>Small area estimation and spatial correlation .....</b>	<b>305</b>
12.1	Setting up the model	306
12.2	Forming the "small area" estimator	312
12.3	The quality of estimators	316
12.4	Implementation with R	320
<b>IV</b>	<b>Part 4: Extensions</b>	
<b>13</b>	<b>Graph partitioning and analysis .....</b>	<b>327</b>
13.1	Graphs and geographical analysis of city networks	328
13.2	Graph partitioning methods	333
<b>14</b>	<b>Confidentiality of spatial data .....</b>	<b>349</b>
14.1	How to evaluate spatial disclosure risk?	351
14.2	How to deal with disclosure risk?	354
14.3	Application for a grid of 1 km <sup>2</sup> squares	360
14.4	Differencing issues	367
	<b>Index .....</b>	<b>375</b>



# Foreword EFGS - EUROSTAT

Mariana Kotzeva - *Director General of Eurostat*

Janusz Dydyczewicz - *President of the European Forum for Geography and Statistics (EFGS),  
Statistics Poland*

The last few years have seen various international and national initiatives to build bridges between the world of geospatial information and that of statistical information.

The UN Economic and Social Council (ECOSOC) on 27 July 2011 recognized the need to promote international cooperation in the field of global geospatial information and therefore set up the Committee of Experts on Global Geospatial Information Management (UNGIM). UN-GGIM adopted decision 3/107 (see E/C.20/2013/17) at its third session, held in the United Kingdom in July 2013, which “acknowledged the critical importance of integrating geospatial information with statistics and socio-economic data and the development of a geospatial statistical framework”.

The Global Statistical Geospatial Framework (GSGF) currently being developed will provide an integrated and interoperable common method for geospatially enabling statistics and managing geospatial information at all stages of statistical production. It connects spatial information that describes our physical man-made and natural environment, and statistics that describe their socio-economic and environmental attributes. This framework has already proven useful for the 2030 Agenda for Sustainable Development and the 2020 Round of Population censuses. Eurostat fully supports these global initiatives and their implementation at the European level. The implementation of the global strategy in Europe relies on methodological guidance developed through the GEOSTAT series of projects. Starting from the very concrete task of representing census data in a European population grid dataset (GEOSTAT 1) the GEOSTAT projects have gradually increased their ambition and scope by developing a model for a point-based geocoding infrastructure for statistics based on address, buildings and/or dwelling registers (GEOSTAT 2) or developing and testing the GSGF in the European context (GEOSTAT 3).

These projects benefitted from the fruitful exchanges of ideas at the annual conferences of the European Forum for Geography and Statistics (EFGS), funded by Eurostat. The main goal of the EFGS is to promote the integration of statistical and geospatial information and the use of geospatial information in decision making. The EFGS is a voluntary body comprising geospatial experts but also statisticians, researchers working in National Statistical Institutes (NSIs) and experts from National Mapping and Cadastral Agencies (NMCAs). Countries outside Europe are also active in the EFGS, which aims to establish a global forum and good cooperation with UN-GGIM at regional and global level. EFGS became the official Observer Organization to UN-GGIM: Europe in 2015. EFGS also acts as a reference group for Eurostat and the project group for the GEOSTAT projects.

Both Eurostat and EFGS acknowledge that integrating statistical and geospatial information relies on strong methodological guidance to ensure the quality and comparability of geospatial statistics. This is why they warmly welcome INSEE’s initiative to compile a handbook of geospatial statistics, based on a point-based statistical information system. They fully endorse the objectives of the handbook, namely to promote, develop and consolidate the use of specific statistical methods available to NSIs only within the framework of a point-based system. These methods, ranging from measuring spatial autocorrelation to drawing a spatially balanced sample, by way of managing confidentiality in a spatial context, fully completely within the scope of NSI activities. This tool, which focuses on practical examples and their implementation in R, will undoubtedly make the statistical production process, and the release and analysis of statistical results, more efficient.

We are convinced that this handbook will be useful for experts in all Statistical Offices and National Mapping Authorities worldwide who wish to know more about using geospatial information in statistics. The handbook will be useful for data producers, users and analysts, both experts and beginners, to identify the opportunities of data integration but also to understand the methodological challenges that the integration of the two at times very different data types brings about. We hope that it will inspire many experts to start using geospatial information in statistical production, and to use geospatial statistics in analysis and decision making.

# INSEE Editorial - reader's guide

Vincent Loonis - *Head of the Geographical Methods and Repositories Division (DMRG), INSEE*  
Marie-Pierre de Bellefon - *Head of the Spatial Analysis Methods Section at the DMRG*

## The rationale for a new spatial analysis manual

Noel Cressie was one of the first to publish a "Handbook of Spatial Statistics" (Cressie 1993a). His work, which is clear and detailed, helps delve into the theory of spatial statistics. However, it does not include a guide on how to use the practical use of these methods. Since the publication, advances in theory and computer sciences have gone hand in hand with an increased supply of geolocated data. Many specialists have in turn written textbooks and other guides to spatial statistics, from the very theoretical Pace et al. 2009, Gelfand et al. 2010 or Anselin 2013 to R software user guides such as Bivand et al. 2008, Brunsdon et al. 2015 as well as works combining theory and practice such as Haining 2003, Schabenberger et al. 2004 or Fischer et al. 2009. Among the French-language works, Zaninetti 2005 describes the theory of spatial statistics, while Caloz et al. 2011 are interested in geostatistics. Within INSEE itself, Jean-Michel Floch presented in 2013 how spatial statistics contribute to the study of socio-economic disparities (Floch 2013) and, in 2015, his reflections on spatial statistics in general.

The purpose of this spatial analysis handbook is to answer the questions faced by research teams at statistical institutes: what use should be made of these new geolocated data sources? In what cases should their spatial dimension be taken into account? How should spatial statistical and econometric methods be applied? In contrast to existing manuals, its teaching principle has been expressly designed according to the issues specific to statistical institutes: the examples of application use data collected by public statistical institutes and the emphasis is placed on practice and the importance of parameter selection. The theoretical foundations are explored in sufficient depth to enable an understanding of the subtleties in the practical implementation of methods, referring readers interested in understanding extensions of a higher technical level to specialised works. While the majority of the chapters present well-documented and frequently used methods, some draw on innovative, recently-published work. Among the topics addressed by the INSEE-EUROSTAT manual are sampling and respect for confidentiality, both of which are important points for NSIs, and yet are explored in very little depth by existing works. A few of the chapters open up on concepts currently used only rarely at INSEE, such as geostatistics.

The panel of authors combines statisticians from different departments of INSEE (Department of Statistical Methodology, Department of Regional Action, Department of Economic Studies and Summary Statistics) and university professors (Universities of Le Mans, Paris-Sud, Guyana, Agrosup and INRA Dijon). The drafting of the manual proved an opportunity to encourage interaction between the public statistical community and academia.

## Handbook outline

In 2008, the Nobel Prize for Economics was awarded to Paul Krugman, the father of new economic geography. This award marks the growing importance of taking spatial phenomena into account. Krugman describes economic geography as "that branch of economics that worries about where things happen in relation to one another" (Krugman 1991). This quote illustrates the approach specific to any spatial analysis study, regardless of its field of application. The analyst starts by

describing the location of the observations, then measures the importance of spatial interactions in order to be able to take these interactions into account using an appropriate model. These three stages match the first three parts of the handbook: Part 1: *Describing geolocated data*; Part 2: *Measuring the importance of spatial effects*; Part 3: *Taking spatial effects into account*.

Location is referenced in a geographic information system using spatial coordinates. One of the characteristics of spatial analysis is therefore that the medium for observation, defined as all spatial coordinates of the objects to be processed, contains potentially meaningful information for the analysis. To make use of such information, the person in charge of the study usually starts by grouping the data according to their geographical proximity. This is the first step before exploring the characteristics of data location and describing the evolution of variables in space. This grouping is also a key parameter for ensuring the confidentiality of the data disseminated by public statistical institutes. The first chapter of the manual — *Descriptive spatial analysis* — explains how data can be taken up using the R software in order to make the first maps. Concepts in cartographic semiology are also introduced. The second stage of spatial analysis consists in defining an object's neighbourhood. Defining the neighbourhood is an essential step toward measuring the strength of spatial relationships between objects, in other words the way in which neighbours influence each other. The endeavour in the second chapter of the manual — *Codifying the neighbourhood structure* — is to succeed in defining neighbourhood relationships consistent with the actual spatial interactions between objects. This chapter introduces several concepts of neighbourhood, based on contiguity or distances between observations. The issue of the weight ascribed to each neighbour is also addressed.

Geolocated data can be divided into three categories: areal data, point data and continuous data. The fundamental difference between these data is not the size of the geographical unit in question, but the process that generated the data. Where areal data are concerned, the location of the observations is assumed to be fixed: it is the value of the observations that follows a random process. For example, the GDP of each region is defined as areal spatial data. The more the observation values are influenced by values of observations that are geographically close to them, the greater the spatial autocorrelation. Spatial autocorrelation indices measure the strength of spatial interactions between observations. The global and local versions of spatial autocorrelation indices are presented in Chapter 3: *Spatial Autocorrelation Indices*. When dealing with point data, the location of the observations is the random variable. This can be, for example, the location of shops within a city. The strength of spatial interactions is therefore measured by the difference between the observations' spatial distribution and a completely random spatial distribution. Chapter 4: *Spatial distributions of points* provides the methods and tools that can be used, for instance, to highlight possible attractions or repulsions between the different types of points and the way in which the significance of the results obtained is assessed. Lastly, continuous data are characterised by the fact that there is a value for the variable of interest at any point in the territory studied. However, these data are measured only in a discrete number of points. This can imply, for example, the chemical composition of the soil that can be used by the mining industry. Chapter 5: *Geostatistics* presents the fundamental concepts by which continuous data can be studied — semi-variogram, interpolation of data using kriging methods, etc.

The third and fourth parts of the manual focus on the study of areal data, which are the most often used in public statistical institutes. The spatial phenomena affecting areal data can be divided into spatial dependence, versus spatial heterogeneity. Spatial dependence means a situation in which the value of an observation is linked to the values of neighbouring observations (either they influence each other or are both subject to the same unobserved phenomenon). Spatial econometrics models this spatial dependency. There are multiple forms of interactions 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. Chapter 6: *Spatial econometrics: common models* details step by step the methodology for choosing a model (estimate and tests), as well as the precautions to be taken in interpreting the results. The way in which spatial econometric models can be applied to the study of panel data is presented in Chapter 7: *Spatial econometrics on panel data*.

Spatial heterogeneity refers to the fact that the influence of explanatory variables on the dependent variable varies with the location of the observations. Geographically weighted regression or spatial smoothing are used to take this phenomenon into account. Regardless of whether a regression model is used, *spatial smoothing* (chapter 8) filters information to reveal the underlying spatial structures. *Geographically-Weighted Regression* (GWR, chapter 9) responds more specifically to the observation that a regression model estimated over the whole of an area of interest may not adequately address local variations. Geographically-Weighted Regression can be used, in particular with the help of associated cartographic representations, to identify where the local coefficients deviate the most from the overall coefficients, and to build tests to assess whether and how the phenomenon is non-stationary.

Whether intended to take spatial dependence or spatial heterogeneity into account, spatial analysis methods have been developed using comprehensive data. However, they can enrich the range of sampling techniques. These techniques are particularly important for public statistical institutes, whose data are often obtained through surveys. Upstream, the choice of the entities to be selected at the first degrees of a sampling plan and the selection of the sample can be improved using the spatial sampling techniques presented in Chapter 10. Downstream, Chapter 11: *Spatial econometrics on survey data* presents the potential pitfalls when estimating a spatial econometric model on sampled data and assesses the potential corrections suggested in empirical literature. Chapter 12: *Estimation on small areas and spatial correlation* presents small area methods and how taking spatial correlation into account can improve estimates.

The fourth part of the manual — *Extensions* — introduces two chapters that directly use the spatial dimension of data, while moving away from the classical treatment of spatial dependence or spatial heterogeneity. Network analysis allows all flows between territories to be taken into account to determine privileged relations. The techniques of *graph analysis and partitioning* are presented in Chapter 13. The profusion of geocoded data goes hand in hand with a high risk of disclosure, as the number of variables needed to uniquely identify a person decreases considerably when the person responsible for the intrusion knows the exact geographical position of an individual. This subject is crucial for statistical institutes, which face high demand for the dissemination of sensitive data at ever-finer geographical levels. Chapter 14: *Confidentiality of spatial data* aims to provide suggestions on how to assess and manage the risk of disclosure, while preserving spatial correlations.

The first three chapters are recommended to all readers, as they make it easier to understand the entire handbook. The foreword to each individual chapter further specifies which chapters should be read in advance to correctly understand the chapter at hand. The body of the text presents the fundamental theory and examples of practical application. The boxes are more technical extensions and are not essential to understand the essence of the method.

**References - INSEE Editorial**

- Anselin, Luc (2013). *Spatial econometrics: methods and models*. Vol. 4. Springer Science & Business Media.
- Bivand, Roger S et al. (2008). *Applied spatial data analysis with R*. Vol. 747248717.
- Brunsdon, Chris and Lex Comber (2015). *An Introduction to R for Spatial Analysis Et Mapping*. Sage London.
- Caloz, Régis and Claude Collet (2011). *Analyse spatiale de l'information géographique*. PPUR Presses polytechniques.
- Cressie, Noel (1993a). *Statistics for spatial data*. John Wiley & Sons.
- Fischer, Manfred M and Arthur Getis (2009). *Handbook of applied spatial analysis: software tools, methods and applications*. Springer Science & Business Media.
- Floch, Jean-Michel (2013). « Détection des disparités socio-économiques, l'apport de la statistique spatiale ».
- Gelfand, Alan E et al. (2010). *Handbook of spatial statistics*. CRC press.
- Haining, Robert P (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.
- Krugman, Paul R (1991). *Geography and trade*. MIT press.
- Pace, R Kelley and JP LeSage (2009). « Introduction to spatial econometrics ». *Boca Raton, FL: Chapman & Hall/CRC*.
- Schabenberger, Oliver and Carol A Gotway (2004). *Statistical methods for spatial data analysis*. CRC press.
- Zaninetti, Jean-Marc (2005). *Statistique spatiale: méthodes et applications géomatiques*. Hermès science publications.

# Authors et reviewers

## **Descriptive spatial analysis**

**Authors:** Sophie Audric, Marie-Pierre de Bellefon, Eric Durieux

**Reviewers:** Maëlle Fontaine, François Sémécurbe

## **Codifying the neighbourhood structure**

**Authors:** Marie-Pierre de Bellefon, Ronan Le Gleut, Vincent Loonis

**Reviewers:** Salima Bouayad-Agha, Ali Hachid

## **Spatial autocorrelation indices**

**Authors:** Marie-Pierre de Bellefon, Salima Bouayad-Agha

**Reviewers:** Olivier Sautory, Lionel Védrine

## **Point configurations**

**Authors:** Jean-Michel Floch, Eric Marcon, Florence Puech

**Reviewers:** Salima Bouayad-Agha, Gabriel Lang

## **Geostatistics**

**Author:** Jean-Michel Floch

**Reviewers:** Marie-Pierre de Bellefon, Thomas Romary

## **Spatial econometrics : common models**

**Authors:** Jean-Michel Floch, Ronan Le Saout

**Reviewers:** Salima Bouayad Agha, Pauline Givord, Julie Le Gallo, Olivier Sautory

## **Spatial econometrics on panel data**

**Authors:** Salima Bouayad-Agha, Julie Le Gallo, Lionel Védrine

**Reviewers:** Sébastien Faivre, Alain Pirotte

## **Spatial smoothing**

**Authors:** Laure Genebes, Auriane Renaud, François Sémécurbe

**Reviewers:** Valérie Darriau, Jean-Michel Floch

## **Geographically Weighted Regression**

**Authors:** Marie-Pierre de Bellefon, Jean-Michel Floch

**Reviewers:** Maëlle Fontaine, François Sémécurbe

## **Spatial sampling**

**Authors:** Cyril Favre-Martinoz, Maëlle Fontaine, Ronan Le Gleut, Vincent Loonis

**Reviewers:** Eric Lesage, Patrick Sillard

## **Spatial econometrics on survey data**

**Authors:** Raphaël Lardeux-Schutz, Thomas Merly-Alpa

**Reviewer:** Ronan Le Saout

**Small area estimation and spatial correlation**

**Authors:** Pascal Ardilly, Paul Bouche, Wencan Zhu

**Reviewers:** Jean-François Beaumont, Olivier Sautory

**Graph analysis and partitioning**

**Authors:** Pascal Eusébio, Jean-Michel Floch, David Lévy

**Reviewers:** Laurent Beauguitte, Benjamin Sakarovitch

**Confidentiality of spatial data**

**Authors:** Maël-Luc Buron, Maëlle Fontaine

**Reviewers:** Maxime Bergeat, Heïdi Koumarianos

**Acknowledgments**

We thank the steering committee of the handbook for helping us to define the orientations to give to this manual, and for proofreading the whole manual :

Salima Bouayad-Agha, Vianney Costemalle, Valérie Darriau, Marie-Pierre de Bellefon, Gaël de Peretti, Sébastien Faivre, Jean-Michel Floch, Maëlle Fontaine, Pauline Givord, Vincent Loonis, Olivier Sautory, François Sémécurbe, Patrick Sillard.

We also thank Hicham Abbas, Jérôme Accardo, Maël-Luc Buron, Julie Djiriguian, Sonia Oujia and Pascale Rouaud for their help in proofreading the handbook's final version.

We thank Kathleen Aubert for her precious help in proofreading the translation.

We thank Brigitte Rols for designing the cover.



# Part 1: Describing geolocalized data

1	Descriptive Spatial Analysis .....	3
2	Codifying the neighbourhood structure	31



# 1. Descriptive Spatial Analysis

SOPHIE AUDRIC, MARIE-PIERRE DE BELLEFON, ERIC DURIEUX  
INSEE

---

<b>1.1</b>	<b>Different types of spatial data</b>	<b>4</b>
1.1.1	Point data . . . . .	4
1.1.2	Continuous data . . . . .	5
1.1.3	Areal data . . . . .	6
<b>1.2</b>	<b>Concepts in cartographic semiology</b>	<b>7</b>
1.2.1	What is cartographic semiology? . . . . .	7
1.2.2	The objectives of a map . . . . .	7
1.2.3	To each type of data, its visual variable . . . . .	7
1.2.4	Some advice . . . . .	8
<b>1.3</b>	<b>Mapping data with R</b>	<b>10</b>
1.3.1	Manipulating spatial objects . . . . .	12
1.3.2	Producing statistical maps . . . . .	18
1.3.3	<i>sf</i> : the future of spatial data processing under R . . . . .	20
1.3.4	From the surface to the point, and vice versa . . . . .	23
<b>1.4</b>	<b>Examples of studies using aggregated spatial data</b>	<b>26</b>
1.4.1	Access to green spaces - Statistics Sweden . . . . .	26
1.4.2	Regional poverty rate - European ESPON programme . . . . .	27
1.4.3	Optimal location of wind turbines - British Cartographic Society . . . . .	29

---

## Abstract

The objective of spatial analysis is to understand and explore the entanglement between the spatial positioning of objects and phenomena and their characteristics. The literature traditionally distinguishes three types of spatial data – point data, continuous data and areal data. To each type of data correspond specific analytical methods. However, whatever the nature of the spatial data, the first step consists in manipulating them and aggregating them at a geographical scale appropriate to the underlying spatial process. Mapping the data can offer a synthesised view of a situation, make it understandable to a broader audience and give insight into which statistical tools would be best-suited to continue the study. This first descriptive analysis can also, when taking place as part of a study, bring to light specific problems in the data (collection, missing data, outliers, etc.) or lead to invalidate certain hypotheses necessary for the development of econometric methods. We have incorporated into this chapter the concepts in cartographic semiology needed to produce a quality map.<sup>1</sup>

---

1. These concepts in semiology were taken from the INSEE publication "Guide to Cartographic Semiology" (2017) and to which a large number of people contributed, whom we thank.

This chapter describes how spatial data can be manipulated using the R software and how the first descriptive maps be created. Studies carried out at various European statistical institutes are used to illustrate these concepts.

## 1.1 Different types of spatial data

Spatial data is an observation we know the value and the location of. The support of observations, defined as the set of spatial coordinates of the objects to be processed, offers a potentially rich source of information for the process analysis.

Some properties of spatial data contradict the assumptions necessary to the use of standard statistical methods. For instance, the hypothesis that observations are independent, a requirement in most econometric models, is not verified when *spatial dependence exists* – when the value of observation  $i$  influences the value of the neighbouring  $j$  observation. Another possible characteristic of spatial data is *spatial heterogeneity*: the influence of explanatory variables on the dependent variable depends on the location in space. A variable may prove influential within one neighbourhood, but not in another. Many methods have thus been specifically developed to analyse spatial data.

The methods and their objectives depend on the nature of the spatial data involved. According to the classification suggested by Cressie 1993b, three types of spatial data can be identified:

- point data;
- continuous data;
- areal data.

The fundamental difference between these data is not the size of the geographical unit in question, but the process that generated the data.

### 1.1.1 Point data

Point spatial data are characterised by the **spatial distribution** of the observations. The data generating process generates the geographic coordinates associated with the emergence of an observation. The value associated with the observation is not studied; only the location counts. The latter can be, for example, the location where a disease emerges during an epidemic, or how certain tree species are distributed in space. Spatial analysis of point data is aimed at **quantifying the gap between the spatial distribution of observations and a completely random distribution in space**. If the data are more aggregated than if they had been randomly distributed across the territory, clusters can be identified and their significance measured.

 The main methods for analysing point data are described in Chapter 4: "Spatial distributions of points"

■ **Example 1.1 — Cluster Detection.** Fotheringham et al. 1996 have focused on detecting significant clusters of uncomfortable houses. They compare the spatial distribution of uncomfortable houses as identified on the ground with the distribution that would have emerged if they were distributed randomly among all houses. The hypotheses on random distribution in space make it possible to assess the significance of house groupings (Figure 1.1). ■

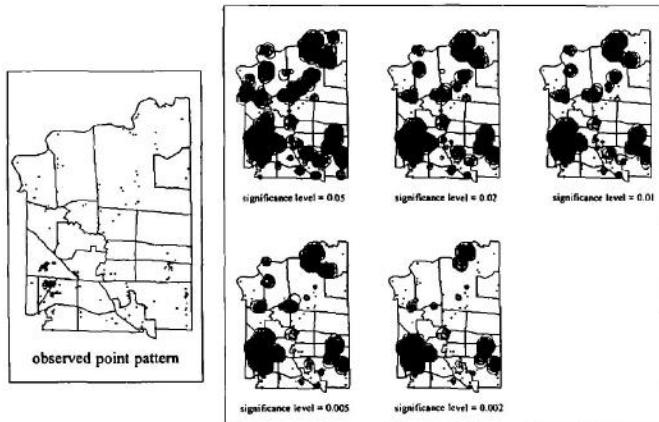


Figure 1.1 – Detecting significant clusters

**Source:** Fotheringham et al. 1996

### 1.1.2 Continuous data

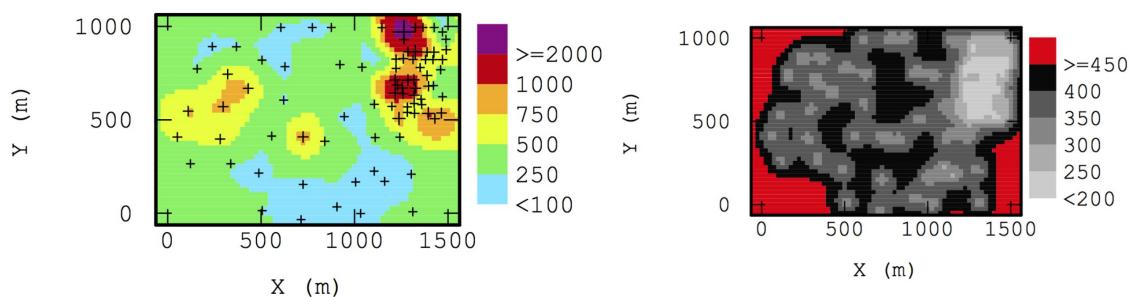
With continuous data, there is a value for the variable of interest at any point across the territory studied. Data are generated on a continuous basis, across a subset of  $\mathbf{R}^2$ . However, these data are measured only in a discrete number of points. These include, for example, the chemical composition of the soil (data beneficial to the mining industry), water or air quality (for studies on pollution), or various meteorological variables. Spatial analysis of continuous data, also referred to as geostatistics, is aimed at predicting the value of a variable at a point where it has not been sampled, as well as the reliability of this prediction. Geostatistics also helps optimise the data sampling plan.



The main methods for analysing continuous data are described in Chapter 5: "Geostatistics".

#### ■ Example 1.2 — Predicting pollution. Chiles et al. 2005

The researchers at the GeoSiPol (Practices of geostatistics in polluted sites) working group take into account the spatial dependency structure between data using the *kriging* technique. They predict the quantity of pollutant found in places where the soil has not been sampled and quantify the estimation uncertainty (Figure 1.2). ■

Figure 1.2 – Predicting the pollutant content in soil (mg/kg/m<sup>2</sup>) (left) and the standard deviation of this prediction (right)

**Source:** GéoSiPol manual - Mines de Paris Chiles et al. 2005

### 1.1.3 Areal data

Where areal data are concerned, while the location of the observations is assumed to be fixed, the associated values are generated according to a statistical process. These data are most often based upon a partition of the territory into contiguous zones, but they can also be fixed points on the territory. This includes, for example, GDP by region, or the number of marriages per town hall. The term 'areal' is therefore misleading, as these data are not necessarily represented on a surface. The focus here is on the relationship between **values of neighbouring observations**. The spatial analysis of areal data begins with **defining the neighbourhood structure of the observations**, then proceeds to **quantify the influence that observations have on their neighbours**, and lastly **assesses the significance of this influence**.

- R The main techniques used for analysing areal data are described in chapter 2: "Codifying neighbourhood structure" and chapter 3: "Spatial Autocorrelation Indices", as well as in Part 3.

■ **Example 1.3 — Local spatial dependency.** Givord et al. 2016 have aimed to answer the question: "Are privileged lower secondary schools always located in a privileged environment?" For this purpose, the authors use *local spatial autocorrelation indices*<sup>2</sup>. These indices compare the similarity between a lower secondary school's social level and that of its environment with the similarity they would have if the same various social levels of lower secondary schools were randomly distributed among the schools. Local spatial autocorrelation indices make it possible to identify the lower secondary schools for which the influence of the surrounding social environment is significant (Figure 1.3).

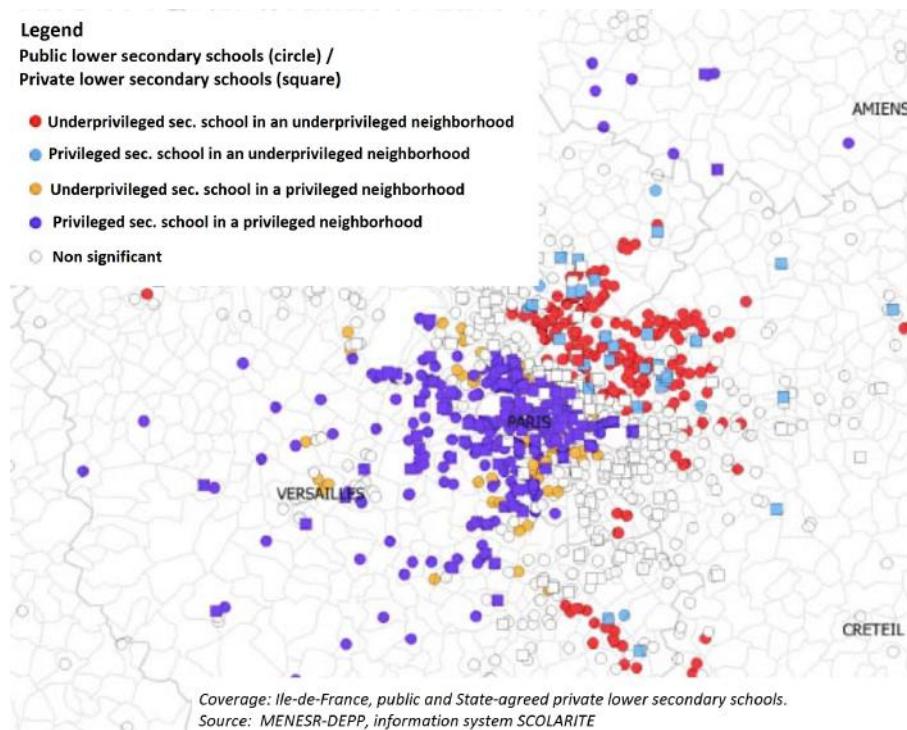


Figure 1.3 – Influence of the social level of the neighbourhood of a lower secondary school on the social level of the school itself

**Source:** Givord et al. 2016

2. Refer to chapter 3 "Spatial autocorrelation indices" for more details

**Box 1.1.1 — Spatial data may fall into multiple categories.** Dividing spatial data into three categories allows the analyst to choose the most appropriate method. However, it should be kept in mind that these categories are permeable and that the decision to analyse phenomena from one point of view stems from the scale of analysis and the very purpose of the study. For example, a house is considered to be a point object when studying significant groupings across space, but can also be seen as areal data when the aim is to identify the spatial correlation between the ages of the houses' inhabitants.

## 1.2 Concepts in cartographic semiology

### 1.2.1 What is cartographic semiology?

Cartographic semiology is the set of rules that make it possible to convey information as clearly as possible thanks to a cartographic image. It is good to have these rules in mind before moving on to designing a map using the R software. Cartographic semiology is a full-fledged language developed to facilitate communication, using graphic tools referred to as visual variables. When properly used, these variables reinforce the message while also making it clearer.

Visual variables include the shape, texture and size of the object to be depicted, its orientation and its colour. The latter may be connected with transparency effects or display a gradient of colours, according to a given scale of values. Dynamics have appeared more recently as a visual variable, with the creation of such outputs as animated maps.

Visual variables are distinctive for their ability to highlight:

- quantities, often represented as proportional circles;
- a hierarchy, representing an ordered set of relative values, for example population densities;
- differences between entities represented, for example industry and tourism;
- similarities, by grouping into a single set various objects reflecting the same theme.

Moreover, when well-combined, multiple visual variables can stress the point.

### 1.2.2 The objectives of a map

Graphs make it possible to directly and comprehensively grasp information and are an advantageous alternative to lengthy tables. This is even truer for maps. Their main interest lies in how they can integrate the spatial dimension, especially when the number of territories is relatively high. For instance, a map makes it possible to pick up on information at a single glance. The spatial dimension embraces all at once the geographical location, proximity to the coast, the mountains, large cities, neighbouring countries, etc., hence the importance of adding geographical benchmarks – neighbouring regions and countries, city names, rivers, thoroughfare, etc. Moreover, maps serve as good communications tools. They are easy to understand because their territory can generally be recognised and they offer a pleasant illustration. Technological advances in mapping tools, which are free of charge and easy to access, now make it easy to produce attractive maps. However, it is important that aesthetics not take precedence over relevance, and even more that they do not distort the information provided by the map.

### 1.2.3 To each type of data, its visual variable

The first question to be settled is that of the knowledge the map is intended to convey. To represent a variable in volume terms or an absolute number, proportional circles are used. For ratios, densities, trends, shares and typology, solid colour maps are recommended. Bilocated data or flows are best illustrated by radial flow maps, proportional arrows or vector resultants. Lastly, the location, for example of equipment, is shown using maps with symbols.

With solid colour maps (also referred to as class analysis or choropleth maps), the positive values are shown in warm tones (red, orange) while negative values generally appear in cool tones (blue, green). Moreover, a hierarchy in values can be reflected using a colour gradient, with the darkest (or brightest) colours reflecting the extreme values.

There are also rules for discretising data, *i.e.*, how observations are grouped into classes. The number of classes is calculated according to the number of observations. There are several theories for determining the optimal number. According to the Sturges rule, for example, it is equal to  $1 + 3 * \log_{10}(N)$ , where  $N$  is the number of observations.

In practice:

- for fewer than 50 observations: 3 classes;
- for 50 to 150 observations: 4 classes;
- for more than 150 observations: 5 classes.

The form of data distribution can also facilitate this choice. For instance, a class is added where both negative and positive values are found. Once the number of classes has been determined, a grouping method has to be chosen. There are several methods for doing so, each with advantages and drawbacks.

- **The quantile method** consists in using the same number of values per class. It results in a harmonious and easy-to-read map, on which the colours in the key are distributed equally. However, it is not always suited to the distribution of data.
- **The "classes of the same amplitude" method:** in this method, the interval between values is divided into ranges of the same length. This method is simple to understand but is very rarely suited to the distribution. Some classes may contain no value at all.
- **The Jenks and Kmeans methods** are designed to create homogeneous classes by maximising the variance between classes and minimising the variance within them. Unlike both previous methods, these methods are perfectly suited to data as they eliminate threshold effects. However, the Jenks method can entail a very protracted calculation time, if a large number of observations needs to be processed. In the latter case, the Kmeans method, which enables quicker calculation time even with a high number of observations, can be used. However, it can be unstable, resulting in different classes for a single dataset. This problem can be managed by repeating the Kmeans multiple times in order to keep the best limits in the end.
- **The manual arrangement method:** in this method, the mapper defines the limits of the classes. It is useful when aiming to highlight significant values (zero or near-zero boundary, average...) or to marginally improve the positioning of certain thresholds in accordance with local distribution. It also makes it possible to make different maps comparable with one another, by setting identical class boundaries. This method requires that the data distribution be analysed in advance, first using the Jenks or Kmeans method to develop homogeneous classes and then manually adjusting the class boundaries to prevent any threshold effects.

#### 1.2.4 Some advice

- **One simple message per map.** Maps are often difficult to understand when they contain too much information. For example, because it is overly complicated, no message emerges from the map shown in picture 1.4. For an effective map, the rule "less is more" applies. Concretely, this means limiting the number of variables to be depicted on the same map.
- **Show the basic information.** A map must contain an informative title (most often connected with a descriptive sub-heading), a reference to the zoning shown, a key, a source and a copyright. The scale, logo or North arrow may also be included.
- **Not depicting the territory as an island.** It is best when readers are also provided with environmental information, as this enables them to locate the territory shown – for example,

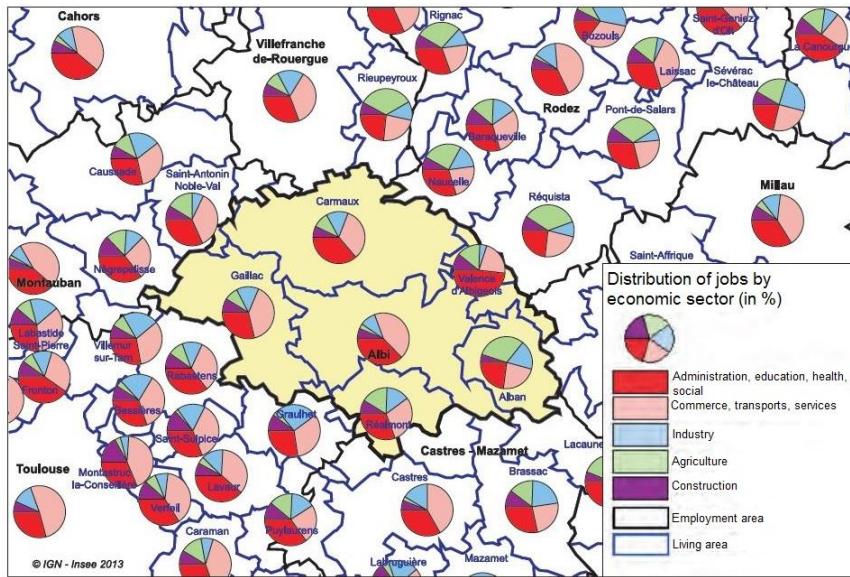


Figure 1.4 – Breakdown of jobs by business sector in living areas

bordering departments or regions, topographic elements such as the sea or the road system. In Figure 1.5, it would have been wise to represent the municipalities of the surrounding departments, in particular Dijon in the north or Lyon in the south, so as to illustrate the title, which is not very explicit.

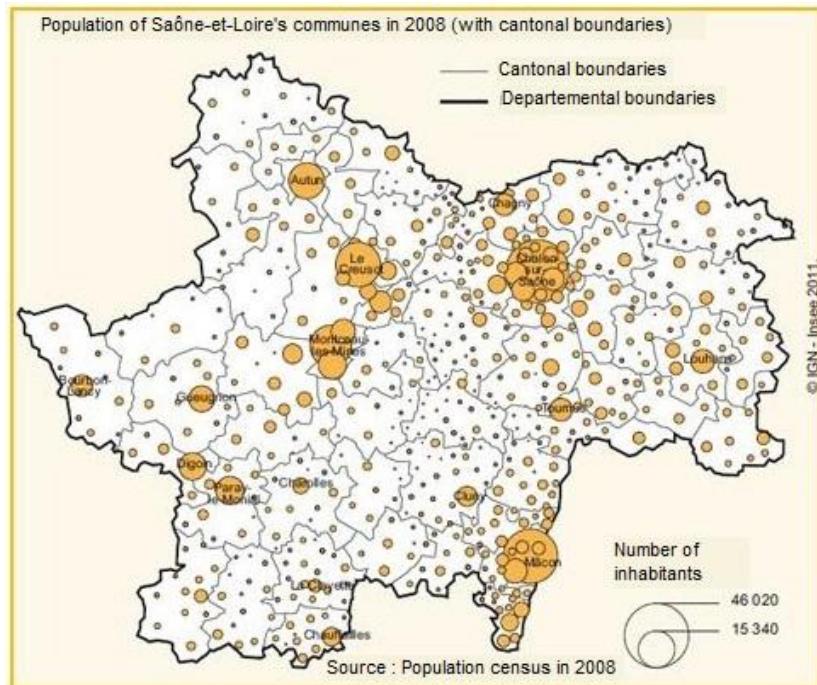


Figure 1.5 – Multiple medium-sized cities

Furthermore, it may be interesting to extend the analyses carried out on the territory to the surrounding environment, on the condition that the territory of interest emerges clearly, as in Figure 1.6 (dark green contour and light green line). The expanded analysis here makes it

possible to identify the geography of Toulouse's demographic dynamism compared to that of Bordeaux and to better understand the importance of the Languedoc urban system, in line with that of the Rhône corridor.

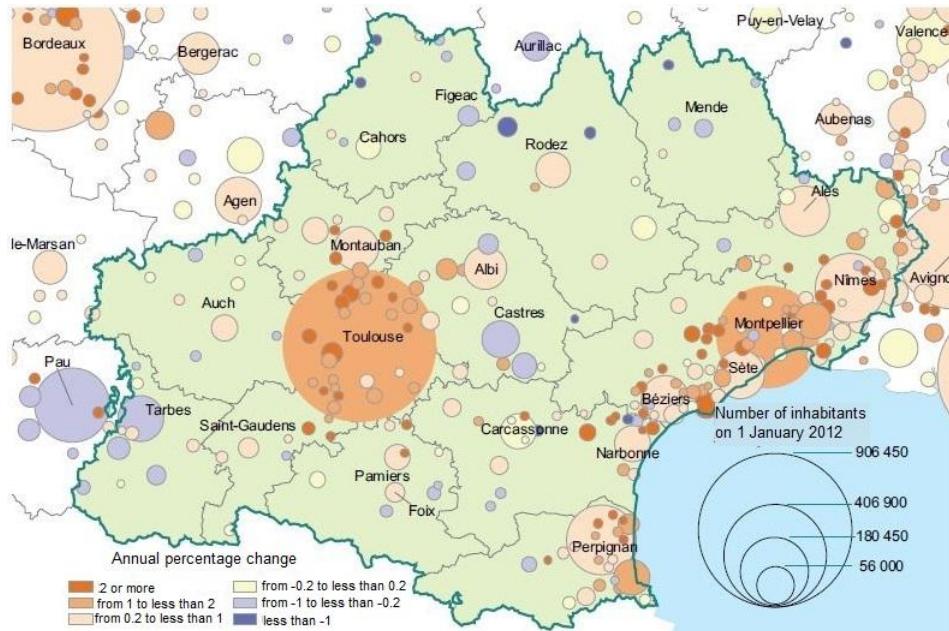
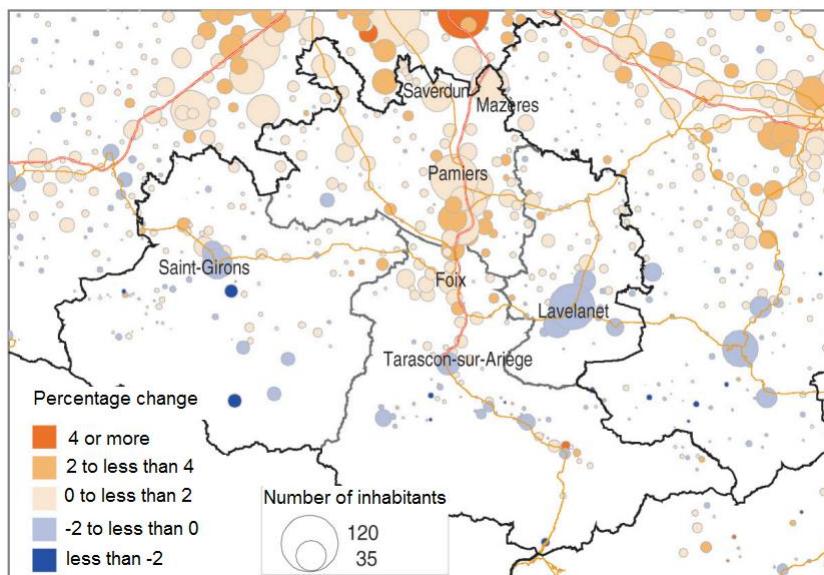


Figure 1.6 – A monocentric urban system around Toulouse and polycentric on the coast

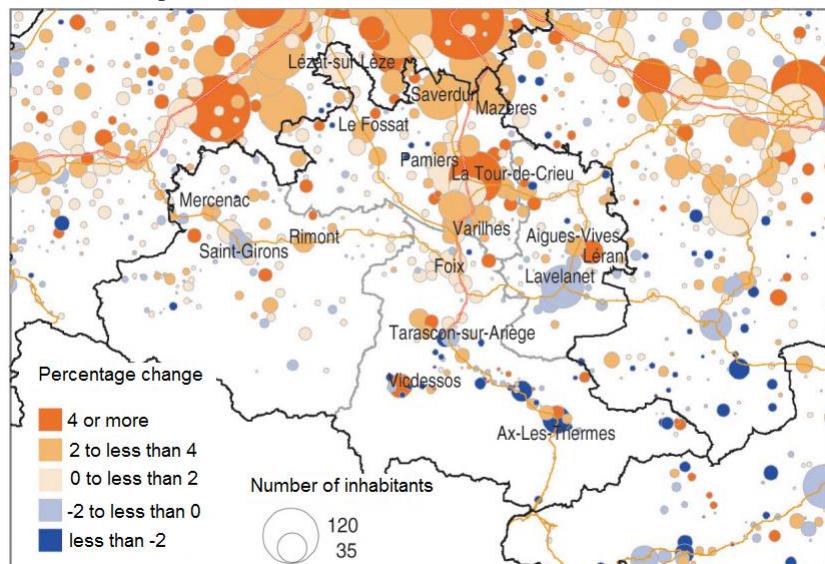
- **Comparable maps.** When two maps showing the same territory with the same visual variables are placed side by side or below one another, it is an incentive for the user to make comparisons. To facilitate that process, both maps should have a harmonised key (same classes, circles or arrows) and the same scale with identical zoom. In the maps on Figure 1.7, the harmonised legends make it possible to compare the annual trend in population over the two periods from 1982-2011 and 2006-2011.
- **Choosing an indicator: proportion or volume?** Class-based analysis is used to represent a sub-population in relative value (or proportion) or a trend. It is to be proscribed when representing numbers of individuals or volumes, as it could lead the reader to misinterpret the map. The eye would establish a correspondence between the volume represented and the surface of the coloured territory. For instance, a class analysis of the number of inhabitants per municipality would lead to a visual overestimation of the population of Arles, the largest municipality in France. Furthermore, class-only analysis can sometimes be misleading as high percentages can sometimes apply to small numbers. This is why it is sometimes necessary to combine this type of analysis with a proportional circle analysis covering the numbers of individuals. Depending on the message we want to convey, we will choose to colour circles with a class analysis (Figure 1.7) or superimpose circles on a class analysis (Figure 1.8). When using coloured circles, the eye is more attracted to the size of the circles, while with the superimposed circles, the eye will first be drawn to the darkest colours in the class analysis.

### 1.3 Mapping data with R

Geolocated data can be aggregated on a more or less large geographical scale. They can then be mapped in different ways. In this section, we will describe how to simply start out in mapping



(a) Average annual trend in the population of the municipalities between 2006 and 2011



(b) Average annual trend in the population of the municipalities between 1982 and 2011

Figure 1.7 – Average annual trend in the population of the municipalities of Basse-Ariège  
**Source:** INSEE, *Population census 1982, 2006, 2011*

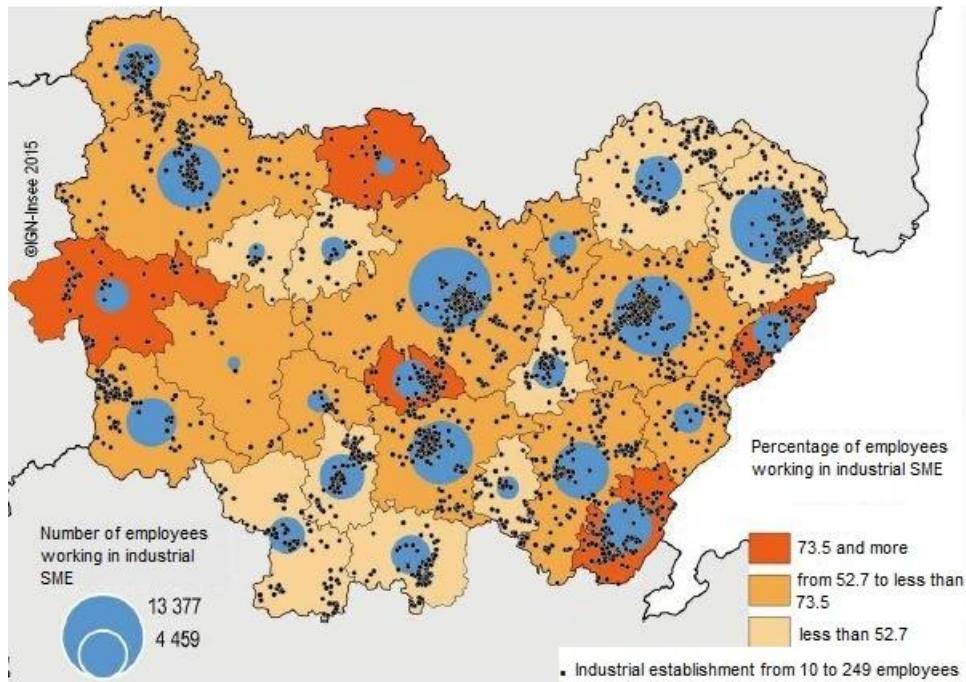


Figure 1.8 – Breakdown of employees working in an SME in the industrial sector

Source : INSEE, Local Knowledge of Productive Resources - CLAP 2012

with R, and present some appropriate packages. Many packages can be used to represent spatial data. The ones we will implement in this manual are:

- *sp*: basic package defining spatial objects;
- *rgdal*: import/export of spatial objects;
- *rgeos*: geometric manipulation;
- *cartography*: producing analysis maps.

We will also present the *sf* package that groups together all the functions of packages *sp*, *rgdal* and *rgeos*.

### 1.3.1 Manipulating spatial objects

#### Points, polygons, lines

The *sp* package makes it possible to create or convert various geometries into an *sp* object such as points, lines, polygons or grids, for instance. In general, each *sp* object is made up of different parts known as slots. Each slot contains specific information (geographical coordinates, table of attributes, system of coordinates, spatial scope, etc.)

Access to a slot for an *sp* object will take place *via* the operator @ (*objet@slot*).

Spatial objects can be addressed in different forms. The first corresponds to **points**, *i.e.* a set of georeferenced points.

---

```
library(sp)
```

```
# contents of a communal table containing the coordinates of the town halls
# in WGS84 (latitude/longitude)

head(infoCom)
##          nom_commune latitude longitude préfecture
```

---

```

##           <chr>    <dbl>    <dbl>    <chr>
## 1 Faches-Thumesnil 50.58333  3.066667  Lille
## 2          Lille 50.63333  3.066667  Lille
## 3      Lezennes 50.61667  3.116667  Lille
## 4          Lille 50.63333  3.066667  Lille
## 5      Ronchin 50.60000  3.100000  Lille
## 6 Villeneuve-d'Ascq 50.68333  3.141667  Lille

# Transforming into a spatial object

municipalities<-  SpatialPoints(coords=infoCom[,c(2,3)])

#Viewing available slots

slotNames(municipalities)
## [1] "coords"      "bbox"        "proj4string"

# Understanding the spatial scope

municipalities@bbox  # ou bbox(municipalities)
##           min       max
##latitude 50.000000 51.083333
##longitude 2.108333 4.183333

```

---

This object can also be represented graphically *via* the standard graphic instruction `plot` (illustration in Figure 1.9).

---

```
plot(municipalities)
```

---

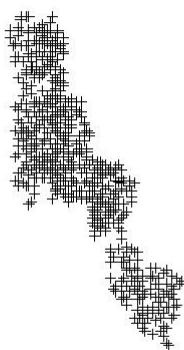


Figure 1.9 – Municipalities of Northern France

**Source:** INSEE

Our spatial object can also have a table of attributes describing the geographical objects it contains. The object then belongs to the `SpatialPointsDataFrame` class:

---

```
#Adding the attribute table
```

---

```
nord <- SpatialPointsDataFrame(coords=infoCom[,c(2,3)], data=infoCom[, c(1,4)])
```

---

This table of attributes is accessed *via* the new slot created @data:

---

```
nord@data
##           nom_commune préfecture
##           <chr>      <chr>
## 1   Faches-Thumesnil    Lille
## 2       Lille    Lille
## 3   Lezennes    Lille
## 4       Lille    Lille
## 5   Ronchin    Lille
## 6 Villeneuve-d'Ascq    Lille
## 7   La Madeleine    Lille
## 8       Lille    Lille
## 9   Comines    Lille
## 10  Deulemont    Lille
## # ... with 611 more rows
```

---

The creation of **georeferenced polygons**, although slightly more complex, follows the same logic.

First of all, we will create simple polygons using the coordinates of the vertices:

---

```
# Creating the sets of coordinates
x1 <- c(439518.5, 433091.8, 455774.1, 476566.1, 476944.2, 459554.4,
       439518.5)
y1 <- c(8045280, 8031293, 8018439, 8026756, 8044902, 8054731, 8045280)
       c1 <- data.frame(x1, y1)
x2 <- c(444929.2, 417667.9, 501837.1, 499792.5, 444929.2)
y2 <- c(8121306, 8078029, 8067465, 8109039, 8121306)
       c2 <- data.frame(x2, y2)
x3 <- c(456530.1, 450481.5, 472785.8, 476566.1, 456530.1)
y3 <- c(8101608, 8089510, 8087620, 8099717, 8101608)
       c3 <- data.frame(x3, y3)

# Creating the polygons
p1 <- Polygon(coords = c1, hole = F)
p2 <- Polygon(coords = c2, hole = F)
p3 <- Polygon(coords = c3, hole = T)
```

---

The parameter `hole` is used to identify polygons representing holes inside other polygons.

These objects have 5 slots, including:

- `@labpt`, which provides the coordinates of the centre;
- `@hole`, which establishes whether it is a hole;
- `@coords`, which allows the coordinates of the vertices to be retrieved.

They can then be assembled into multiple polygons:

---

```
P1 <- Polygons(sr1 = list(p1), ID = "PolygA")
P2 <- Polygons(sr1 = list(p2, p3), ID = "PolygB")
```

---

Consequently, polygon  $P1$  will be composed of  $p1$  and  $P2$  will be  $p2$  with a hole in the centre defined by  $p3$ .

They still have 5 different slots, including:

- `@Polygons` which lists the polygons used for its creation;
- `@ID` which provides the identifiers given to the polygon.

We then spatialise this set of polygons to make it a single spatial object:

---

```
SP <- SpatialPolygons(Sr1 = list(P1, P2))
```

---

Our spatial object is structured as follows: the `SpatialPolygons` contains a list of two polygons (multiple polygons), each containing a list of `Polygons` (single polygons), which contain the coordinates that delineate them. Thus, to access the coordinates of the first simple polygon contained in the second multiple polygon, we have to write:

---

```
SP@polygons[[2]]@Polygons[[1]]@coords
```

---

```
##           x2      y2
## [1,] 444929 8121306
## [2,] 499793 8109039
## [3,] 501837 8067465
## [4,] 417668 8078029
## [5,] 444929 8121306
```

---

To add an attribute table to our geographical object, simply create a `dataframe` containing as many lines as there are multiple polygons in our object. The lines must be sorted in the same order as the polygons and each line identified by the same identifier.

---

```
Info <- c("Simple", "Hole")
Value <- c(342, 123)
mat <- data.frame(Info, Value)
rownames(mat) <- c("PolygA", "PolygB")

SPDF <- SpatialPolygonsDataFrame(Sr = SP, data = mat)
```

---

A new `@data` slot is added to retrieve the table of attributes. This object can be represented graphically, resulting in Figure 1.10.

---

```
plot(SPDF, col=c("lightgrey", "black"))
```

---

Objects such as **georeferenced lines** can be constructed in the same way as that shown previously for the polygons. This time, the functions `SpatialLines` and `SpatialLinesDataFrame` will be used.

Thus introduced, our municipal, departmental, etc. background maps will be objects of `SpatialPolygons` (`DataFrame`) type, our road or waterway background maps will be of `SpatialLines` (`DataFrame`) type and our airport or townhall background maps will be of `SpatialPoints` (`DataFrame`) type.

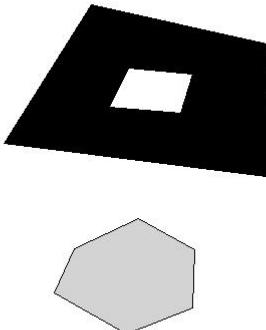


Figure 1.10 – Polygons generated

### Working on a vector layer

Most of the time, we do not create geographical objects from scratch, and instead manipulate objects that already exist. Multiple packages can be used to import or export geographical objects. The simplest and most complete remains *rgdal* which makes it possible to read and manipulate a very large number of formats. The most common vector format is the "ESRI ShapeFile", which provides a base of maps through 5 files to be shown side-by-side in the same folder (.shp, .shx, .dbf, .prj, .cpg). All these files have the same name; only the extension will differ.

To import the background map, the function `readOGR` is used:

---

```
library(rgdal)
comr59<- readOGR(dsn = "My_path\\", layer = "comr59", verbose = FALSE)
```

---

The parameters of function `readOGR` are:

- `dsn`: the pathway to the folder containing the files;
- `layer`: file name (without extension).

The result is then an object of R-type `SpatialPolygonsDataFrame` (example in Figure 1.11).

---

```
class(comr59)
## [1] "SpatialPolygonsDataFrame"
## attr(,"package")
## [1] "sp"
plot(comr59)
```

---

`readOGR` makes it possible to import a wide range of cartographic formats. When using a background map from MapInfo, the syntax changes little:

---

```
comr59MI<- readOGR(dsn= "My_path\\comr59.tab", layer="comr59", verbose=
FALSE)
```

---

As with the ShapeFiles, the MapInfo format is composed of a number of files that must all be found in the same folder, with the same name but different extensions. In this case, the `dsn` points to the `.tab` file, and the `layer` takes the name of the files in the background map.

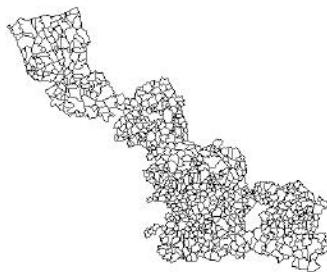


Figure 1.11 – North municipal background map

**Source:** INSEE

To select a subset of our map, refer to the associated dataframe *via* slot @data. For instance, to select municipalities with a surface area exceeding  $200 \text{ km}^2$ :

---

```
comr59_extended <- comr59[comr59@data$surf_m2>20000000, ]
```

---

To view this selection, the 2 objects are superimposed by colouring the selection in grey (the result can be seen in Figure 1.12):

---

```
plot(comr59)
plot(comr59_extended,col ="darkgrey",add =TRUE)
```

---

Parameter add=TRUE makes it possible to superimpose the 2 funds.

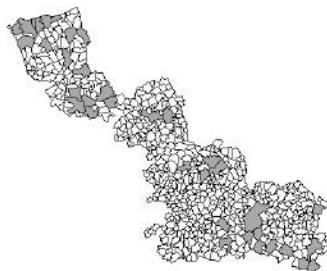


Figure 1.12 – Extended North municipal background map

**Source:** INSEE

To save our new cartographic background map, we use the `writeOGR` function which has as parameters:

- `obj`: R object to be exported;
- `dsn`: backup folder path;
- `layer`: common name of files (without extension);
- `driver`: object export format.

All possible formats are provided by the `ogrDrivers()` function.

For instance, to export our selection in the ShapeFile format:

---

```
writeOGR(comr59_extended,dsn="My_path",layer="comr59_extended",
driver="ESRI Shapefile")
```

---

In MapInfo format:

---

```
writeOGR(comr59_extended, dsn="My_path\\comr59_extended_MI.tab",
layer="comr59_extended_MI", driver="MapInfo File")
```

---

### 1.3.2 Producing statistical maps

#### Projection system

Spatial data are always associated with a projection system. The latter is identifiable by the slot `@proj4string`.

---

```
comr59@proj4string
## CRS arguments:
## +proj=latlong +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0
```

---

A projection system can be assigned to a created object. Let's use again previous presented polygons. They do not by default have any related projection system. To signify that they are in WGS84:

---

```
SPDF@proj4string <- CRS("+proj=latlong +datum=WGS84 +ellps=WGS84")
```

---

The EPSG standard now makes it possible to identify projection systems with a single code. It is 4326 for WGS84. In this context, the previous assignment could be coded:

---

```
SPDF@proj4string <- CRS("+init=epsg:4326")
```

---

All EPSG code matches can be obtained by executing `make_EPSG()`. Thus for Lambert 93 (used inter alia by IGN) the EPSG code is 2154.

To change the projection of a geographic object into a new coordinate system, the `spTransform()` function is used

---

```
comr59_193 <- spTransform(comr59, CRSobj=CRS("+init=epsg:2154"))
```

---

This projection anew is necessary in particular to be able to superimpose two background maps that do not have the same coordinate system.

To produce maps very simply, we present the package *cartography* which in addition to being easy to handle is relatively comprehensive in its possibilities.

#### Maps in proportional symbols:

Mapping of stock data (such as population, number of equipment...) is done using symbols proportional to the size represented. The most common is the circle, but any other symbol can also be used. The code below makes it possible to call up Figure 1.13.

---

```
library(rgdal)
library(cartography)
metr_nice <- readOGR(dsn="My_path",layer="metr_nice",verbose=F)

# Population data table
head(donnees_communales)
```

---

```

##   CODGEO           LIBGEO REG DEP P13_POP
## 1 01001 L'Abergement-Clémenciat 84 01    767
## 2 01002 L'Abergement-de-Varey 84 01    236
## 3 01004 Ambérieu-en-Bugey 84 01 14359
## 4 01005 Ambérieux-en-Dombes 84 01 1635
## 5 01006 Ambléon 84 01    108
## 6 01007 Ambronay 84 01 2503

#Background map plot
plot(metr_nice)

#addition of analysis
propSymbolsLayer(spdf=metr_nice,df =donnees_communales, spdfid = "Codgeo",
                 dfid = "CODGEO",var = "P13_POP",col ="salmon",
                 symbols="circle",legend .pos="right")

#map presentation
layoutLayer(title= "Population of Nice Cote d'Azur metropolis",
            author = "INSEE", sources = "Census 2013",
            scale = NULL, north = TRUE)

```

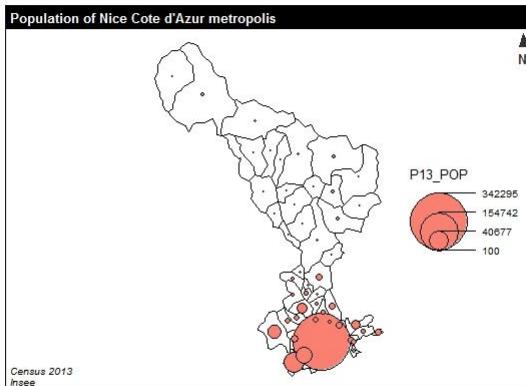


Figure 1.13 – Proportional symbols

**Source:** INSEE, 2013 census

The parameters of the function are:

- spdf: The SpatialPolygonsDataFrame;
- df: the dataframe containing the data to be analysed;
- spdfid: the map mesh identifier (in the slot @data);
- dfid: the line identifier in the dataframe. Must match the former;
- var: the dataframe variable to be analysed.

Other parameters exist and can be listed using the function.

#### **Choropleth maps:**

To depict the rates, solid colour or choropleth maps are used. The variable is divided into classes

and a colour gradient reflects the increase of values.

---

```
plot(metr_nice)

choroLayer(spdf=metr_nice, df =donnes_communales4, spdfid = "Codelgeo",
dfid = "CODGEO", var = "TCHOM", nclass=4, method="fisher-jenks",
legend.pos="right")

layoutLayer(title= "Unemployment rate at municipality level in the
metropolis of Nice Cote d'Azur",
author = "INSEE", sources = "Census 2013",
scale = NULL, north = TRUE)
```

---

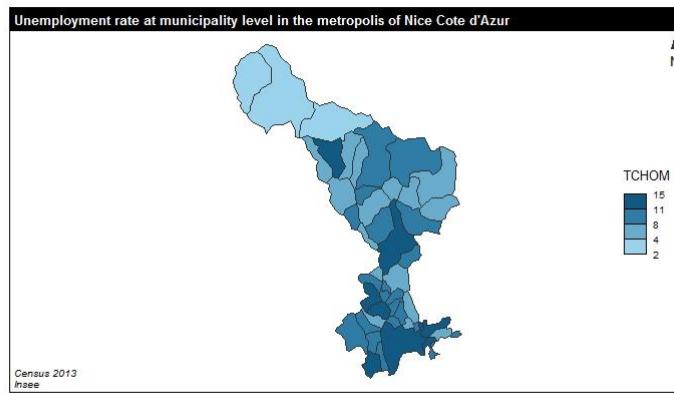


Figure 1.14 – Choropleth maps

**Source:** INSEE, 2013 census

The classification is done either by specifying the number of classes (`nclass`) and the grouping method (`method` that makes it possible to choose from among the methods shown in section 1.2) or by providing a threshold vector (`breaks`).

#### Other mapping functions:

- `propSymbolsChoroLayer`: this is a combination of proportional symbols and choropleth maps (to simultaneously represent a number of unemployed workers and an unemployment rate, for example);
- `typoLayer`: to represent a typology by specifying a qualitative variable and a colour vector of the same length as the number of modalities;
- `gradLinkLayer`: to represent flows or links.

Other packages make it possible to produce statistical maps using R. One example is:

- `RgoogleMaps`: To produce maps using road rasters or satellite GoogleMaps;
- `leaflet`: To produce interactive maps with OpenStreetMap Raster that can be inserted into web pages or even RShiny.

### 1.3.3 **sf: the future of spatial data processing under R**

As we have seen previously, cartographic data has been processed up to now using three main R packages:

- *sp* to implement spatial type classes;
- *rgdal* for input/output libraries;
- *rgeos* for operations on geometric objects.

Most recently, there has been a single package, named *sf*, which brings together all the functionalities of these 3 packages combined together. It provides users with a unique class for handling all spatial objects. In this chapter, we quickly present the main features of package *sf*. To better understand this package, handling rich geometry or managing inputs/outputs, the reader may refer to the various vignettes made available with the package on the CRAN website. This package is not compatible yet with all the spatial analysis packages presented in this manual, which are most often constructed using *sp*, *rgdal* and *rgeos*.

It is notable that inputs/outputs are much faster with *sf* than with *rgdal*.

Since objects of class *sf* are defined as *data.frame* augmented with geometric attributes, the manipulation of geographical objects is simplified, and is natively made, just like what is done in R for any table. Concretely, the package defines three classes of different objects:

- *sf*: a *data.frame* with spatial attributes;
- *sfc*: the column of the *data.frame* storing geometric data;
- *sfg*: the geometry of each recording.

A spatial object will thus be represented as shown in Figure 1.15.

simple feature collection with 96 features and 4 fields						
geometry type: MULTIPOLYGON						
dimension: XYZ						
bbox: xmin: 99225.97 ymin: 6049647 xmax: 1242375 ymax: 7110480						
epsg (SRID): NA						
proj4string: +proj=lcc +lat_1=44 +lat_2=49.00000000001 +lat_0=46.5 +lon_0=3 +x_0=700000 +y_0=6600000 +datum=NAD83 +units=m +no_defs						
First 10 features:						
CODGEO	LIBGEO	INTREG	REG	geometry		
1 01	Ain	ES	82	MULTIPOLYGON Z (((943513 65...		
2 02	Aisne	NE	22	MULTIPOLYGON Z (((790281 69...		sf
3 03	Allier	ES	83	MULTIPOLYGON Z (((77281 65...		
4 04	Alpes-de-Haute-Provence	SE	93	MULTIPOLYGON Z (((1016633 6...		
5 05	Hauts-Alpes	SE	93	MULTIPOLYGON Z (((1022838 6...		
6 06	Alpes-Maritimes	SE	93	MULTIPOLYGON Z (((1077507 6...		
7 07	Ardèche	ES	82	MULTIPOLYGON Z (((848816 64...		
8 08	Ardennes	NE	21	MULTIPOLYGON Z (((873032.1 ...		
9 09	Ariège	SO	73	MULTIPOLYGON Z (((632344 61...		
10 10	Aube	NE	21	MULTIPOLYGON Z (((838365 67...		

Figure 1.15 – Representation of a spatial object with the *sf* package

Importing existing map background is simplified under *sf* and is formatted as follows:

---

```
library(sf)
depf<- st_read("J:/CARTES/METRO/An15/Shape/Depf_region.shp")
```

---

It should be noted that there is no need to specify the import driver. *st\_read* automatically adapts to the input file format. The function is compatible with the vast majority of common cartographic formats (ESRI-Shapefile, MapInfo, PostGIS, etc.). Spatial data can easily be mapped with the *plot* function (Figures 1.16 and 1.17).

Background map export is just as simple:

---

```
st_write(depf, "U:/fond_dep.shp")
```

---

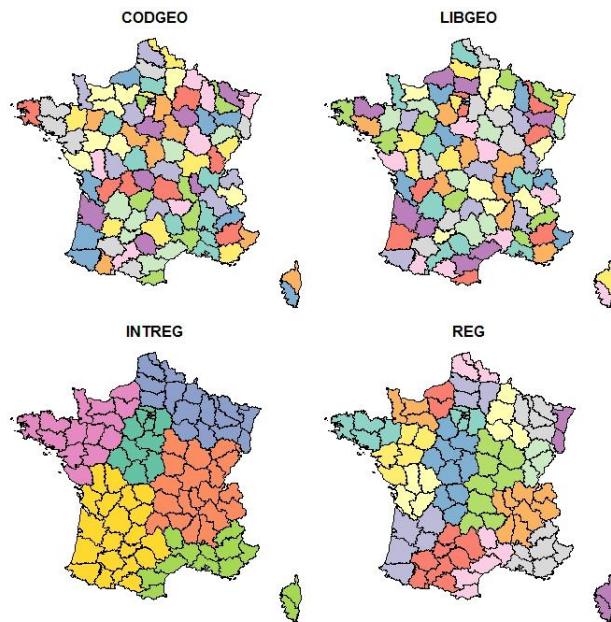


Figure 1.16 – Map obtained with code: `plot(depf)`

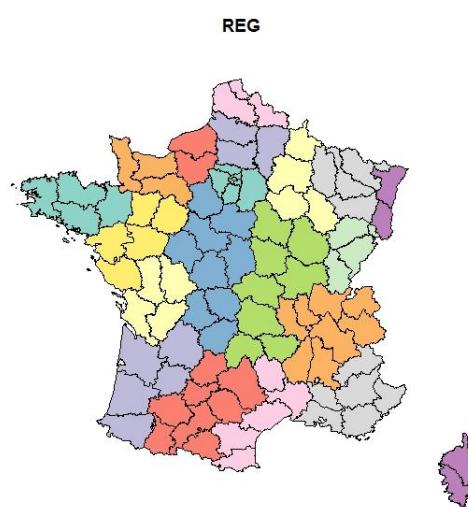


Figure 1.17 – Map obtained with code : `plot(depf ["REG"])`

More generally, package *sf* offers a set of spatial data operators, all bearing the *st\_* prefix and presented in Figure 1.18. The *sf* package is also fully integrated into the *tidyverse* environment

```
## [1] "st_agr<-.sf"           "st_agr.sf"          "st_as_sf.sf"
## [4] "st_bbox(sf"            "st_boundary.sf"    "st_buffer.sf"
## [7] "st_cast.sf"             "st_centroid.sf"     "st_convex_hull.sf"
## [10] "st_coordinates.sf"     "st_crs<-.sf"        "st_crs.sf"
## [13] "st_difference.sf"       "st_geometry<-.sf"   "st_geometry.sf"
## [16] "st_intersection.sf"     "st_is.sf"           "st_line_merge.sf"
## [19] "st_make_valid.sf"       "st_polygonize.sf"   "st_precision.sf"
## [22] "st_segmentize.sf"       "st_set_precision.sf" "st_simplify.sf"
## [25] "st_split.sf"            "st_sym_difference.sf" "st_transform.sf"
## [28] "st_triangulate.sf"      "st_union.sf"         "st_voronoi.sf"
## [31] "st_zm.sf"
```

Figure 1.18 – All the operators found in package *sf*

and thus perfectly matches the functionality of package *dplyr*:

---

```
library(dplyr)
depf<- left_join(depf,pop_dep,by="CODEGEO")
```

---

or also the code below, illustrated in Figure 1.19.

---

```
library(dplyr)
depf %>%
  mutate(
    area = st_area(.), # creating the new variable across the surface area
  ) %>%
  group_by(REG) %>%
  summarise(mean_area = mean(area)) %>%
  plot
```

---

Most packages related to the processing of geographical data have adapted to this new category of objects. Some, like *spdep*, are in the test phase of their adaptation and therefore still require the use of the *sp* package.

Regarding *cartography*, the adjustment has been effective since version 2.0 and the syntax has changed:

---

```
choroLayer(x, spdf, spdfid, df , dfid, var , ...)
```

---

where *x* is an object of the *sf* type. If filled in, the objects *spdf*, *spdfid*, *df* and *dfid* are ignored because all the related information is included in the *x* object.

#### 1.3.4 From the surface to the point, and vice versa

One special feature of areal data is that it may consist in a partition of the whole territory or a set of reference points with distinct geographical coordinates. However, it is easy to move from one representation to another:

- Voronoï polygons create a partition of the territory based on the reference points;
- using the centroid of an area makes it possible to move from a partition of the territory to a set of points.

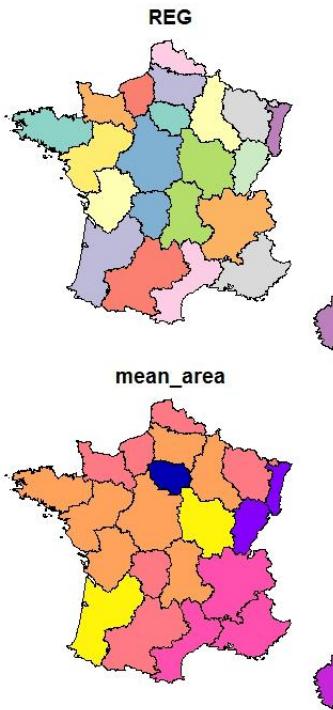


Figure 1.19 – Map produced using packages *sf* et *dplyr*

**Definition 1.3.1 — Voronoï polygon associated with point  $x_i$ .** This is the area of space that is closer to  $x_i$  than to any other point in the set being studied  $\mathbf{x}$ :

$$C(x_i|\mathbf{x}) = \left\{ u \in \mathbb{R}^2 : \|u - x_i\| = \min_j \|u - x_j\| \right\} \quad (1.1)$$

**Box 1.3.1 — Very frequently-used polygons.** Voronoï polygons are among the most widely used geometric structures in the scientific community. According to Aurenhammer 1991, there are three main reasons explaining this interest. The first is that Voronoï polygons are directly observable in nature (in crystalline arrangements, for example). Secondly, they are one of the most fundamental structures defined by a discrete set of points: they show a very large number of mathematical properties and are connected to multiple other fundamental geometric structures. Lastly, Voronoï polygons make it possible to simplify a large number of algorithmic problems. The Voronoï polygon associated with a point is often considered as its "area of influence".

Historically, Gauss (in 1840) and later Dirichlet (in 1850) used Voronoï polygons in their study on quadratic forms. Voronoï took their work one step further to higher dimensions in 1908. A few years later, in 1934, Delaunay built a triangulation associated with the Voronoï polygons and demonstrated the richness of its mathematical properties.

The R package *deldir* makes it possible to calculate the Voronoï polygons associated with a set of points. The *deldir* function returns an object that can be represented with the *plot* function. The package also calculates multiple statistics associated with polygons, such as the surface of each polygon, or the number of its vertices (see detailed documentation).

Many algorithms can be used to build Voronoï polygons (the most effective is the Fortune algorithm) (Fortune 1987). The algorithm implemented by the *deldir* function begins by building a Delaunay triangulation from reference points. This triangulation maximises the triangles' minimal

angle. The vertices of the Voronoï diagram are the centres of the circles circumscribed in the triangles from the Delaunay triangulation. The edges of the Voronoï diagram are on the mediators of the edges of the Delaunay triangulation (the algorithm is detailed in Lee et al. 1980).

### Application with R

---

```
#Packages required
library(deldir)
library(sp)

# Generating random points
x <- rnorm(20, 0, 1.5)
y <- rnorm(20, 0, 1)

#"Deldir" function used to calculate the Voronoï polygons
#based on two sets of geographic coordinates
vtess <- deldir(x, y)

#creates a working window
plot(x, y, type="n", asp=1)

#represents the points
points(x, y, pch=20, col="red", cex=1)

#represents the associated Voronoï polygons
plot(vtess, wlines="tess", wpoints="none", number=FALSE, add=TRUE, lty=1)
```

---

To move from a partition of the territory to a set of points, we can calculate the centroids of the surfaces (Figure 1.20).

**Definition 1.3.2 — Centroid of an S surface.** Point that minimises the average quadratic distance to all S points:

$$\min_c \frac{1}{a(S)} \int_S ||x - c||^2 dx$$

$$c = \frac{1}{a(S)} \int_S x dx$$

Coordinates of  $c$ : average of coordinates of **all S points**

### Application with R

---

```
#Calculating polygon centroids
#From a "Spatial Polygon Data Frame" file

library(GISTools)
centroids <- getSpPPolygonsLabptSlots(polygon)

plot(polygon)
points(centroids, pch = 20, col = "Green", cex=0.5)
```

---

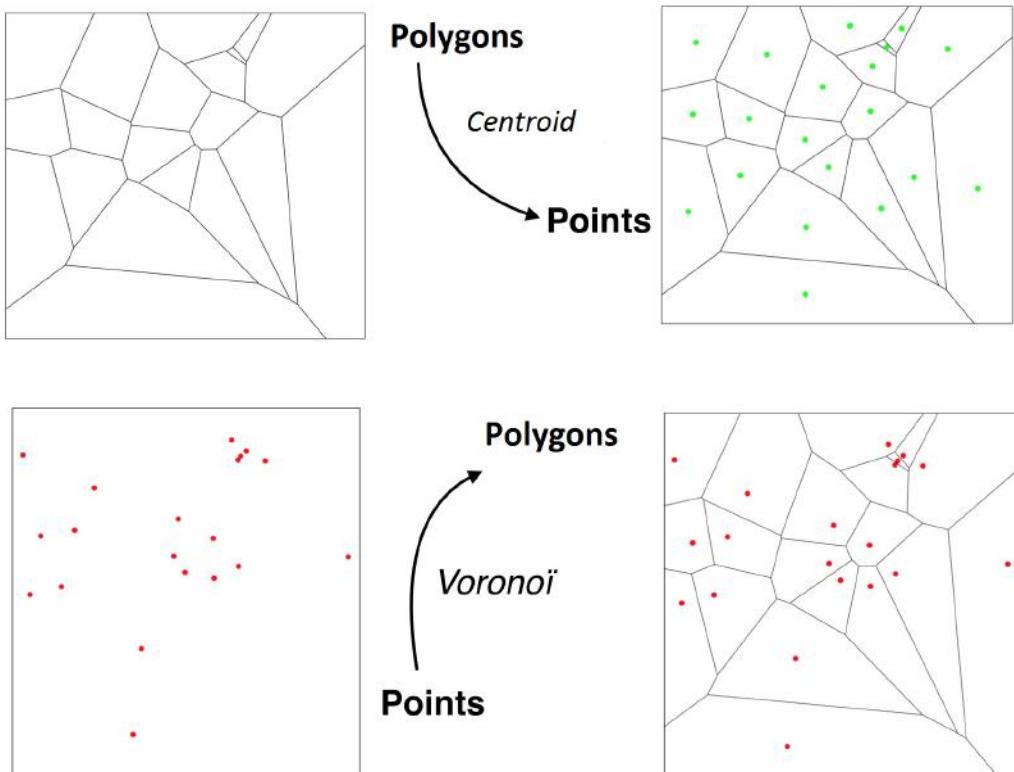


Figure 1.20 – Converting points to polygons and polygons to centroids

## 1.4 Examples of studies using aggregated spatial data

The European Data Integration Group<sup>3</sup> emphasises that representing data on a map with good spatial and temporal resolution makes it possible to detect phenomena that are otherwise invisible. An appropriate representation makes it possible to properly understand the economic, social or environmental situation and to implement relevant public policies. Through the work carried out by three European Statistical Institutes, this section illustrates the variety of descriptive analyses using spatial data – a European project to study regional poverty rates; the analysis of distance to green spaces by the Swedish Statistical Institute; the analysis of optimal location of wind turbines by the British Cartographic Society.

### 1.4.1 Access to green spaces - Statistics Sweden

Increasing access to public green spaces is one of the environmental objectives of the Swedish public policy. In many Swedish municipalities, debate opposes those in favour of increasing the concentration of living spaces with those in favour of preserving green spaces.

The combination of satellite mapping data and localised statistical information from the census makes it possible to better understand the situation on the ground and thus adjust public policies. This study is part of the United Nations' 11th Sustainable Development Goal: "Make cities and human settlements inclusive, safe, resilient and sustainable".

In 2013, the Swedish Statistical Institute drew upon the joint analysis of satellite images and administrative data to characterise the green spaces in Sweden, according to their ownership status and the quality of their vegetation. In most Swedish urban areas, more than 50% of the land is covered by green spaces. On average, three quarters of these spaces are public. Lidingö is the

3. UN-GGIM: United Nations Committee of Experts on Global Geospatial Information Management - Working Group B - Europe

Swedish city covered by the highest proportion of green areas, since they represent approximately 72 % of its total area (figure 1.21) The second part of the study focuses on the accessibility of these green spaces. Using population census data, the Swedish Institute studied the proportion of adults and children living less than a certain distance from a public green space. It found, for instance, that **in 26 Swedish urban areas, less than one percent of the population lives more than 300 meters away from an accessible green space**. In some cities, however, such as Malmö, 15% of children under the age of 6 do not have access to a green space within less than 200 meters from their home (Figure 1.21).

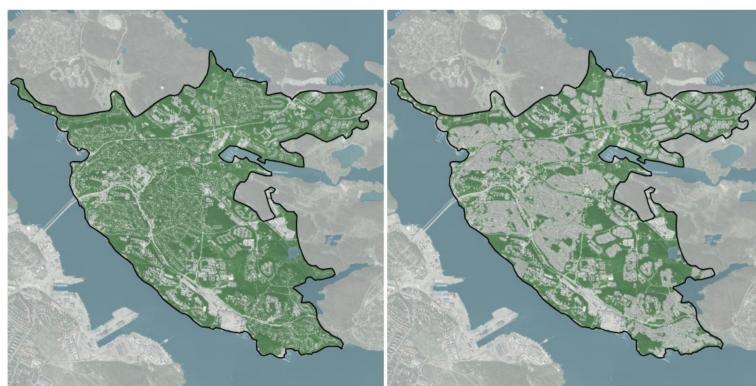


Figure 1.21 – City of Lidingö. Left: all green spaces; Right: green spaces accessible to the public (non-private)

**Source:** Swedish National Institute of Statistics

#### 1.4.2 Regional poverty rate - European ESPON programme

The European EPSON project aims to promote the harmonisation of European public policies by making available regional statistics relevant to decision-makers. Differences in wealth between regions can exacerbate feelings of exclusion and tensions at the national level. Mapping the population's regional poverty rate makes it possible to distinguish the most fragile areas and thus to better target development aid policies.

The poverty threshold (*At-Risk-of-Poverty (ARoP) threshold*) is defined as 60% of the median national standard of living. The poverty threshold therefore varies by country (from €20,362 in Switzerland to €5,520 in Greece). The poverty rate (*ARoP rate*) is defined as the share of individuals whose standard of living is below the national poverty threshold. Figure 1.22 represents the ratio between this indicator calculated at the infra-national level (NUTS3) and the national poverty rate. This makes it possible to **identify the countries with the largest regional disparities**, and to **view the most extreme areas within each country**. The greatest inter-regional disparities in at-risk populations are observed in Turkey, Albania, Hungary, Germany, Croatia, Italy and Spain. The Scandinavian countries, the Netherlands, the Baltic States, Portugal and Greece have a more uniform distribution of *ARoP rates*. Within countries, there are low levels of poverty on the outskirts of capitals and cities, but not necessarily in the cities themselves. The poverty rate is higher in the least accessible regions, such as southern Italy, central Spain or eastern Hungary.

The mapping of poverty rates thus defined helps public decision-making at both national and European levels. To this end, the ESPON programme has published numerous mapping analyses of demographic and social data – a map of male-female ratios by region; various innovation profiles; variations in employment rates or the potential impact of climate change (<https://www.espon.eu/tools-maps>).

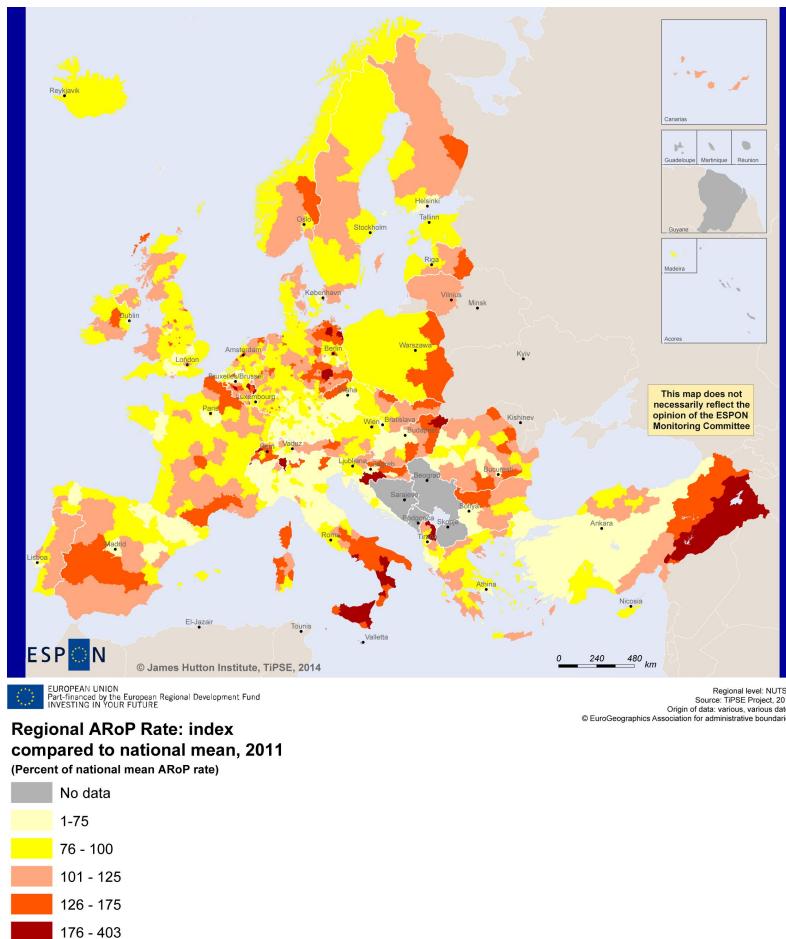


Figure 1.22 – High At Risk of Poverty Indicator

Source: *ESPON Project*

### 1.4.3 Optimal location of wind turbines - British Cartographic Society

The Scottish Government aims to increase renewable energy production by 2020. The Regional Council plays a key role in preserving the local balance, seeking to develop wind farms, while preserving the inhabitants' quality of living. The spatial data provided by the British Cartographic Society (*Ordnance Survey*) have proved very valuable to local decision making.

The objectives of the study are to provide clear and practical guidelines for the location of wind farms. It is planned that the study will take into account many environmental and social factors, such as landscape characteristics and the extent to which views are scenic. The mapping data must be sufficiently detailed to be used by local planners, while remaining easy enough to read and to be understood quickly by all stakeholders (Figure 1.23).

To achieve these objectives, the *Ordnance Survey* worked with many local experts and used many geolocated bases. The various players could monitor the progress of the study using an interactive map. Kevin Belton, GIS Officer, member of the Regional General Council, emphasises the study's added value: "Communicating complex planning information through spatial data has allowed the Council to engage with a wide range of stakeholders, from interested members of the public to commercial developers. This guidance is the end result and it helps ensure developers do not waste resources on applications that are contrary to policy – safeguarding protected areas, the environment and local communities."

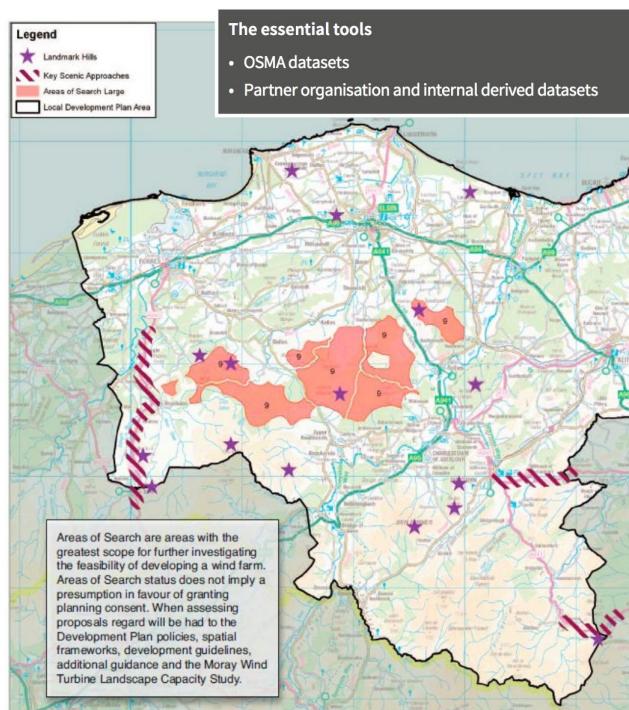


Figure 1.23 – Wind plant implantation study

**Source:** British Ordnance Survey

## References - Chapter 1

- Aurenhammer, Franz (1991). « Voronoi diagrams: a survey of a fundamental geometric data structure ». *ACM Computing Surveys (CSUR)* 23.3, pp. 345–405.
- Bivand, Roger S et al. (2008). *Applied spatial data analysis with R*. Vol. 747248717.
- Chiles, Jean-Paul et al. (2005). *Les pratiques de la géostatistique dans le domaine des sites et sols pollués*. GeoSiPol.
- Cressie, Noel A.C. (1993b). « Statistics for spatial data: Wiley series in probability and statistics ». *Wiley-Interscience, New York* 15, pp. 105–209.
- Fortune, Steven (1987). « A sweepline algorithm for Voronoi diagrams ». *Algorithmica* 2.1-4, p. 153.
- Fotheringham, A. Stewart and F. Benjamin Zhan (1996). « A comparison of three exploratory methods for cluster detection in spatial point patterns ». *Geographical analysis* 28.3, pp. 200–218.
- 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*.
- Lee, Der-Tsai and Bruce J Schachter (1980). « Two algorithms for constructing a Delaunay triangulation ». *International Journal of Computer & Information Sciences* 9.3, pp. 219–242.

## 2. Codifying the neighbourhood structure

MARIE-PIERRE DE BELLEFON, VINCENT LOONIS, RONAN LE GLEUT  
INSEE

---

<b>2.1</b>	<b>Defining neighbours</b>	<b>32</b>
2.1.1	Characteristics of the relationships between spatial objects . . . . .	32
2.1.2	Defining neighbours based on distance . . . . .	34
2.1.3	Defining neighbours based on contiguity . . . . .	39
2.1.4	Defining neighbours based on the optimisation of a trajectory . . . . .	40
<b>2.2</b>	<b>Attributing weights to neighbours</b>	<b>42</b>
2.2.1	From a list of neighbours to a weight matrix . . . . .	42
2.2.2	Importance of the choice of weight matrix . . . . .	45

---

### Abstract

Once the data aggregation scale has been selected and an initial descriptive analysis using mapping tools has been made, the second step of a spatial analysis consists in defining an object's neighbourhood. Defining the neighbourhood is an essential step toward measuring the strength of the spatial relationships between objects, in other words the way in which neighbours influence each other. This makes it possible to compute spatial autocorrelation indices, implement spatial econometrics techniques, study the spatial distribution of observations, as well as perform spatial sampling or graph partitioning.

The challenge in this chapter is to succeed in defining neighbourhood relationships consistent with the actual spatial interactions between objects. This chapter introduces several concepts of neighbourhood, based on contiguity or distances between observations. The issue of the weight assigned to each neighbour is also addressed. Practical implementation is based on R packages *spdep*, *tripack*, *spsurvey* and *tsp*.



Prior reading of Chapter 1: "Descriptive spatial analysis" is recommended.

## 2.1 Defining neighbours

### 2.1.1 Characteristics of the relationships between spatial objects

Consider a surface  $\mathfrak{R}$ . This surface may be divided into  $n$  mutually exclusive zones. Two adjacent zones are separated by a common boundary. Boundaries can arise from spatial discontinuities (administrative or environmental boundaries). They may also rely on Voronoï polygons calculated from points of interest (see Chapter 1: "Descriptive spatial analysis").

**Box 2.1.1 — Mathematical definition of spatial relationships .** Spatial relationships

$\mathcal{B}$  are a subset of the Cartesian product  $\mathbb{R}^2 \times \mathbb{R}^2 = \{(i, j) : i \in \mathbb{R}^2, j \in \mathbb{R}^2\}$  of couples  $(i, j)$  of spatial objects, *i.e.* all couples  $(i, j)$  such that  $i$  and  $j$  are both spatial objects identified by their geographical coordinates, and such that  $(i, j)$  is different from  $(j, i)$ .

A spatial object cannot be linked to itself:  $(i, i) \notin \mathcal{B}$ . Moreover, if  $(i, j) \subseteq \mathcal{B}$  and  $(j, i) \subseteq \mathcal{B}$  for all couples of spatial objects, the spatial relationships are said to be *symmetrical* (Tiefelsdorf 1998).

Spatial relationships are multidirectional and multilateral. They are distinct, in this sense, from temporal relationships, which allow only sequential relationships along the past-present-future axis.

Figure 2.1 illustrates the codifying process of spatial relationships. This approach makes it possible to systematically transcribe the complexity of geographic space into a final set of data analysable by a computer.

First, the study zone is divided into mutually exclusive areas. Each area contains a reference point (often its centroid). Then, the spatial relationships can be specified by a neighbourhood graph connecting the areas considered to be neighbouring, or by a matrix containing the geographical coordinates of the reference points. The third step consists in coding the graph in a neighbourhood matrix, or transforming the geographic coordinates into a distance matrix.

The neighbourhood matrix measures how similar observations are. A value strictly greater than zero indicates that the observations are considered to be neighbouring. For example, in the case of the binary matrix shown in Figure 2.1:

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are spatially linked to each other} \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

Conversely, the distance matrix measures dissimilarity between zones. The higher  $d_{ij}$ , the more different the zones. With, if an Euclidian distance is used:  $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ ,  $\alpha$  and  $\beta$  being the geographical coordinates of the observations.

The neighbourhood matrix is used in the study of areal spatial data, while the distance matrix is rather used for geostatistics (see Chapter 5: "Geostatistics"). However it is possible to move from one to the other by setting a minimum distance beyond which the observations are no longer considered as neighbouring.

The spatial dependence structure may not be geographical. Any relevant dual relationship may be used to define a neighbourhood graph. For instance:

- **at individual level:** friendship bonds, frequency of communication, citations;
- **at company level:** head office-subsidiary ties, similarities in terms of markets;
- **at international level:** strategic alliances, trade flows, shared belonging to an organisation, cultural exchanges and migratory flows.

The following sections detail different neighbourhood specifications.

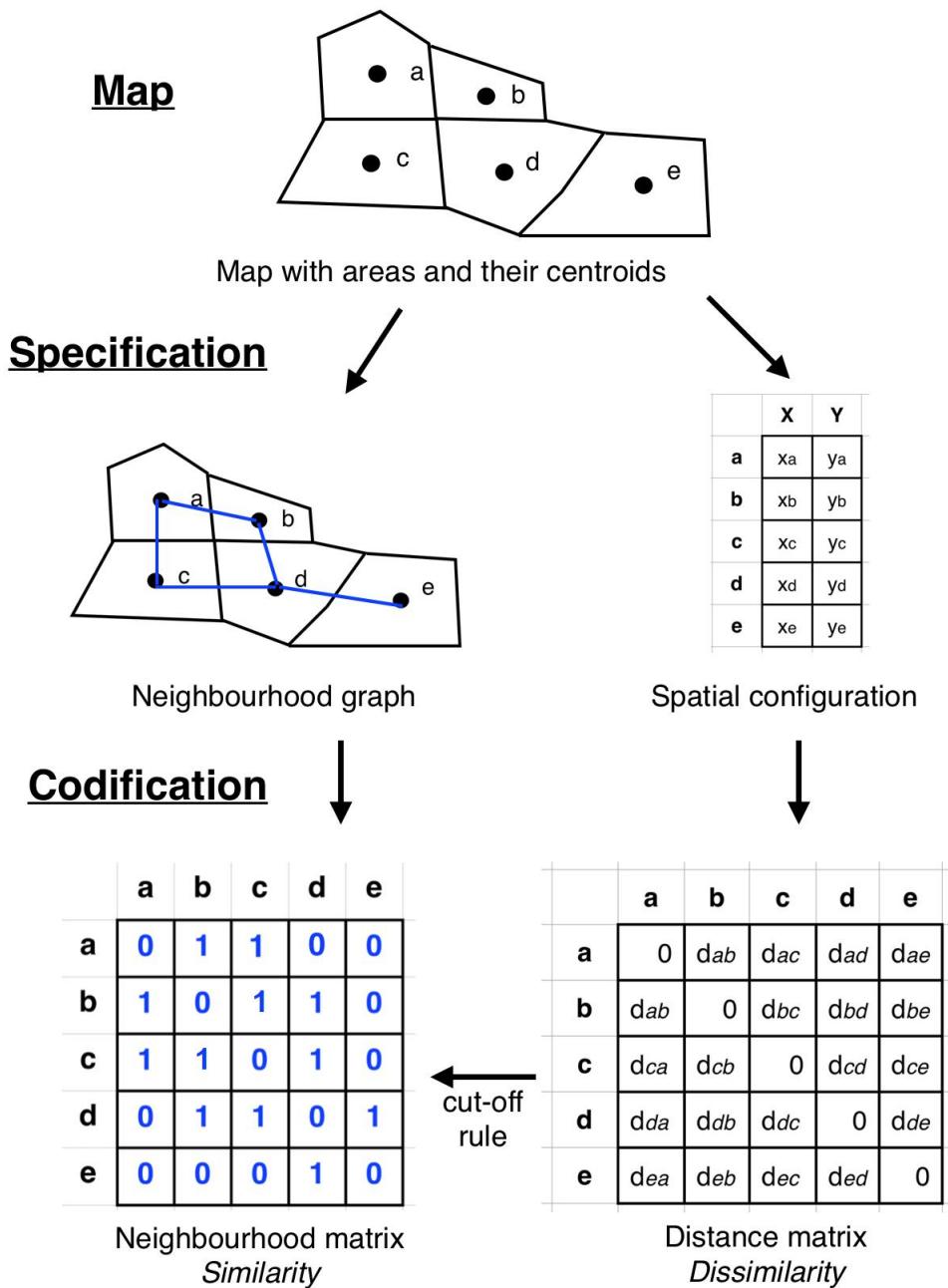


Figure 2.1 – Codifying spatial relations  
Source: Tiefelsdorf 1998

### The "list of neighbours" object in R

Package *spdep* makes it possible to define the relationships between spatial objects. In R, the class of an object defines all its properties and how the statistician can use it. Neighbourhood relationships are recorded in an object of class nb.

Assume  $n$  spatial observations and *neighbours\_nb* the spatial object containing the associated neighbourhood relationships. *neighbours\_nb* is a list of length  $n$ . Each element  $[i]$  of the list contains a vector with the index of the neighbours of the item indexed  $i$ . If  $[i]$  does not have neighbours, the list contains only 0. The list also contains a vector of characters corresponding to the attributes of each neighbourhood zone, as well as a logical value indicating whether the relationship is symmetrical (see Figure 2.2). The main information about the object *neighbours\_nb* can be derived using the function:

```
summary(neighbours_nb)
```

The documentation for package *spdep* provides more information (Bivand et al. 2013b).

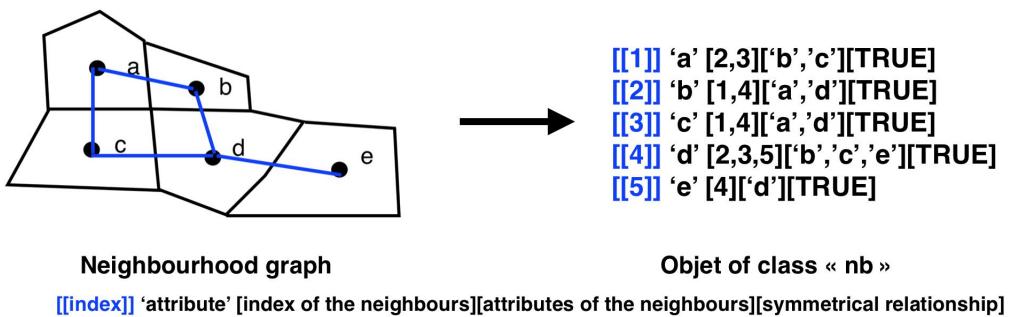


Figure 2.2 – The list of neighbours in *spdep*

### 2.1.2 Defining neighbours based on distance

Once we have a set of points spread across the territory, we can calculate the distance between them. These points may be specific locations where the information has been observed, or points representative of each zone, for example their centroid. In this case, the underlying assumption is that the distribution of the variable within each zone is sufficiently homogeneous to approximate it to a single point.

Neighbourhood graphs materialise the links between the various entities. They are defined in such a way that they represent the underlying spatial structure as closely as possible. There exists many different types of neighbourhood graphs. Here we will show the graphs based on geometric concepts and closest neighbours.

#### Neighbourhood graphs based on geometric concepts

**Delaunay's Triangulation** is a geometric method that connects points into triangles such that the minimum angle of all triangles is maximised (this triangulation is aimed at avoiding "elongated" triangles), see Figure 2.3 and . Delaunay's Triangulation has interesting geometric and mathematical properties. However, the concept of neighbourhood can be refined.

The **sphere-of-influence based graph** links two points if their "circles from the nearest neighbour" overlap. The "circle of the nearest neighbour" of point P is the largest circle centred in P and that contains no other points than P (see Figure 2.4 and 2.5b). The graphs of the sphere of influence are not necessarily connected, *i.e.* all points in the study set are not necessarily interconnected.

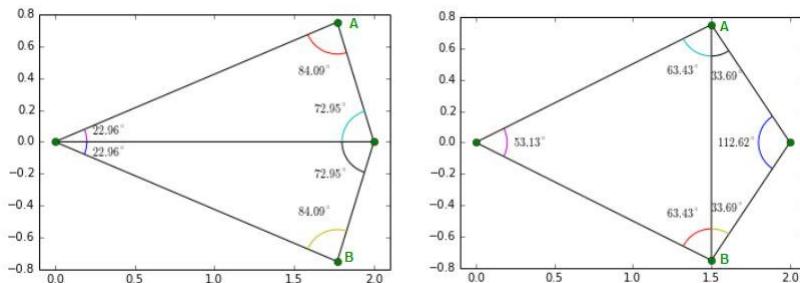


Figure 2.3 – Delaunay triangulation associated with different positions of points A and B

**Source:** Gustavo [CC BY-SA 3.0 (<https://creativecommons.org/licenses/by-sa/3.0>)], from Wikimedia Commons

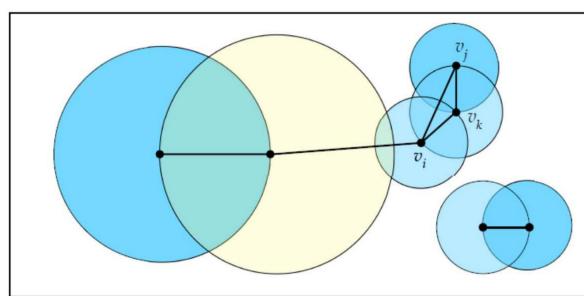


Figure 2.4 – The graph of the sphere of influence of a set of points

**Source:** Toussaint 2014

**Gabriel's graph** links two points  $p_i$  and  $p_j$  if and only if all other points are outside the circle with diameter  $[p_i, p_j]$ . Gabriel's graph removes some links of Delaunay's graph, see Figure 2.5c.

The **graph of relative neighbours** considers that two points  $p_i$  and  $p_j$  are neighbours if

$$d(p_i, p_j) \leq \max[d(p_i, p_k), d(p_j, p_k)] \quad \forall k = 1, \dots, n \quad k \neq i, j \quad (2.2)$$

with  $d(p_i, p_j)$  the distance between  $p_i$  and  $p_j$ . The graph of relative neighbours imposes fewer connections than Delaunay's Triangulation or the sphere of influence graph, see Figure 2.5d. Toussaint 1980 explains that it adapts better to data by requiring the fewest links.

The neighbourhood graphs shown here on Parisian districts are all sub-graphs of Delaunay's Triangulation (see Figure 2.5). They have the advantage that it leaves no unit without neighbours. However, they are only implemented in R with Euclidean distance, while other types of distances, such as the great-circle distance, can be better-suited to certain studies.

### Application with R

---

```

library(rgdal) #To import MIF/MID files
library(maptools) #To import files Shapefile
library(tripack) #To calculate neighbours based on distance
library(spdep)

#Spatial File Import
arr75 <- readOGR("~/ArmF.TAB", "ArmF")

#Neighbours based on the concept of graph
#The input file is a matrix with geographical coordinates
#or an object from type SpatialPoints
coords <- coordinates(arr75)
IDs <- row.names(as(arr75, "data.frame"))

#Delaunay Triangulation
Sy4_nb <- tri2nb(coords, row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy4_nb, coordinates(arr75), add=TRUE, col='red')

#Sphere-of-influence based graph
Sy5_nb <- graph2nb(soi.graph(Sy4_nb, coords), row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy5_nb, coordinates(arr75), add=TRUE, col='red')

#Gabriel Graph
Sy6_nb <- graph2nb(gabrielneigh(coords), row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy6_nb, coordinates(arr75), add=TRUE, col='red')

#Relative neighbours graph
Sy7_nb <- graph2nb(relativeneigh(coords), row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy7_nb, coordinates(arr75), add=TRUE, col='red')

```

---

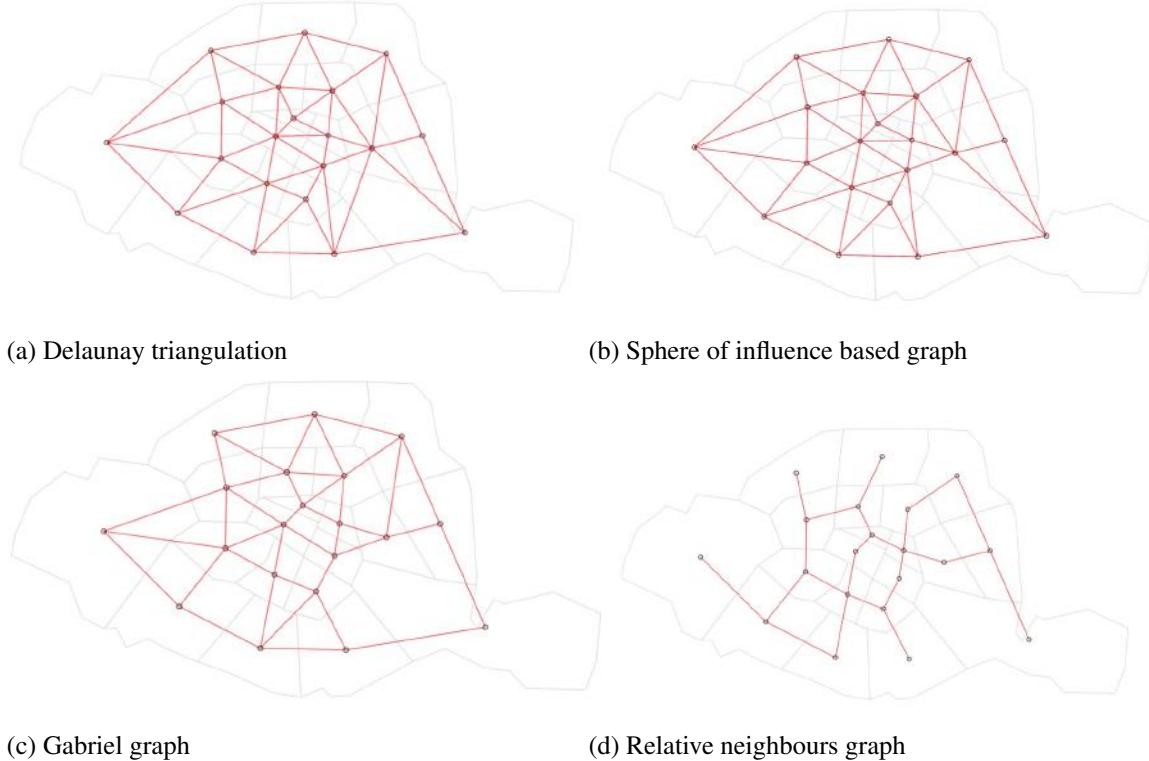


Figure 2.5 – Four neighbourhood graphs of Parisian districts based on geometric concepts

### Neighbourhood graphs based on nearest neighbours

A second method consists in selecting the  $k$  closest points as neighbours (Figure 2.6). This method has the advantage that it leaves no point without a neighbour, which is not required when conducting a spatial analysis, but generally offers a better reflection of reality (a geographical zone is rarely completely isolated). However, it is sometimes difficult to identify the value of  $k$  that reflects the true underlying spatial relationships. The graphs based on the  $k$  closest neighbours are not necessarily symmetrical.

The choice can also be made to keep only the points located at a certain distance. The `nbdists` function of R can be used to calculate the vector of distances between neighbours. It makes it possible to determine the minimum distance  $d_{min}$  above which all points have at least one neighbour, then the `dneareighb` function allows to keep as neighbours only the points between distances 0 and  $d_{min}$ . This "minimum distance" method is not adapted to irregularly spaced data, as the minimum distance required for a relatively isolated point having at least one neighbour is much higher than the distance to the closest neighbour of a point located in a dense zone. There will therefore be significant disparities in the number of neighbours, see Figure 2.6d (Bivand et al. 2013b).

#### Application with R - Source: *Bivand et al. 2013b*

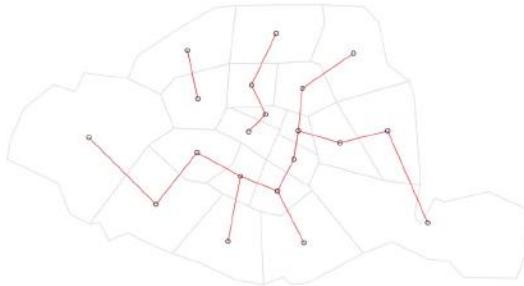
```
#graphs based on the nearest neighbours
Sy8_nb<- knn2nb(knearneigh(coords,k=1),row.names=IDs)
Sy9_nb<- knn2nb(knearneigh(coords,k=2),row.names=IDs)
Sy10_nb<- knn2nb(knearneigh(coords,k=3),row.names=IDs)

plot(arr75, border='lightgray')
plot(Sy8_nb, coordinates(arr75), add=TRUE, col='red')
```

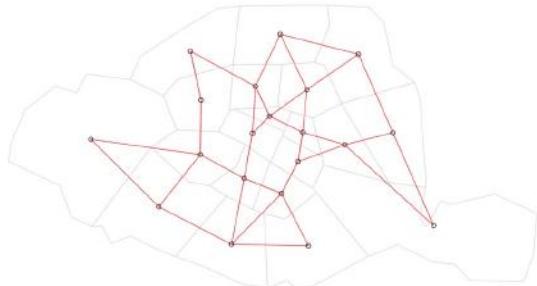
```
#Study of the average distance of the nearest neighbours
dsts <- unlist(nbdists(Sy8_nb,coords))
summary(dsts)
##   Min. 1st Qu. Median    Mean 3rd Qu.    Max.
##   820    1188   1678   1707   2016   3412
max_inn <- max(dsts)

#Calculation and representation of neighbours at the minimum distance
Sy11_nb<- dnearneigh( coords, d1=0, d2=max_inn, row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy11_nb,coordinates(arr75),add=TRUE,col='red')
```

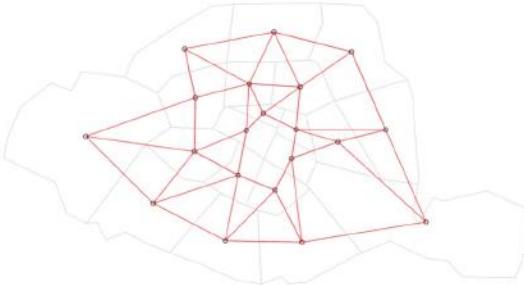
---



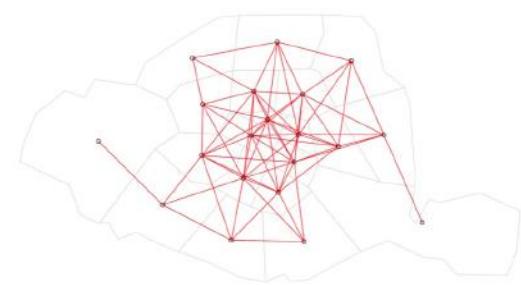
(a) Nearest neighbour



(b) Two nearest neighbours



(c) Three nearest neighbours



(d) Neighbours at a minimum distance

Figure 2.6 – Four graphs based on the nearest neighbours of Parisian districts

### 2.1.3 Defining neighbours based on contiguity

When the areal data consist in a partition of the entire territory, the concept of "distance between observations" can become quite ambiguous. Example 2.1 illustrates the limits of using the distance between centroids to define the notion of neighbourhood.

■ **Example 2.1 — Ambiguity of the notion of distance between centroids.** Let  $R_1, R_2, R_3$  be three distinct zones. It can be considered that since  $R_2$  and  $R_3$  are separated in space, but both are adjacent to  $R_1$ , they are both closer to  $R_1$  than to one another. However, the centroids in these zones are equidistant from each other (see Figure 2.7). Summarising the proximity between zones by the distance between the centroids results in a partial loss of the richness of the spatial relationships.

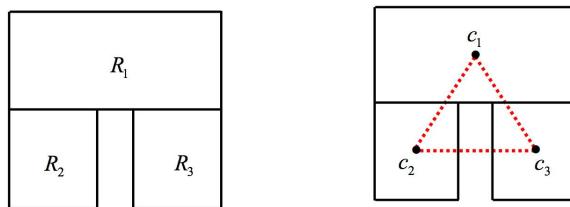


Figure 2.7 – Left: three zones - Right: distance between centroids

**Source:** Smith 2016

This subsection introduces various concepts of contiguity and presents the way in which package *spdep* in R makes it possible to create a list of neighbours.

In the sense of **Rook** contiguity, neighbours have at least two common boundary points (a segment). This matches the movement of the Rook in chess. For two zones to be adjacent in the sense of **Queen** contiguity, they only need to share one common boundary point. This matches the movement of the Queen in chess. Figure 2.8 illustrates these concepts in the case of a regular grid of points. When polygons have an irregular shape and surface, the differences between the Rook and Queen neighbourhoods become more difficult to grasp. It should also be noted that a very large zone surrounded by smaller zones will have a far greater number of neighbours than its neighbouring zones.

The neighbourhood in the sense of contiguity is often used to study demographic and social data, in which it may be more important to be on either side of an administrative boundary than to be located at a certain distance from one another.

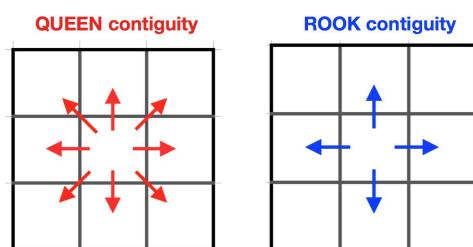


Figure 2.8 – Definition of Queen and Rook contiguity

#### Application with R

Construction of Queen and Rook neighbourhood graphs for Paris districts (Figure 2.9)

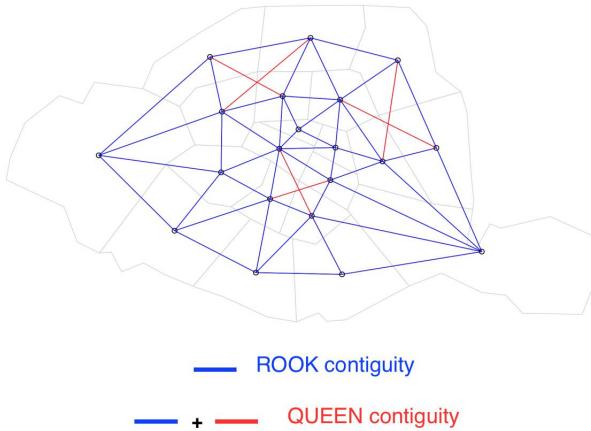


Figure 2.9 – Queen and Rook contiguity in Paris districts

---

```
#The input file is a SpatialPolygons file
#Extraction of list of neighbours as defined in QUEEN contiguity (by
# default)
arr75.nb<- poly2nb(arr75)

#Extraction of list of neighbours as defined in ROOK contiguity
arr75.nb.ROOK<- poly2nb(arr75, queen=FALSE)

#Visual representation of neighbours:
plot(arr75, border='lightgray')
plot(arr75.nb, coordinates(arr75), add=TRUE, col='red')
plot(arr75.nb.ROOK, coordinates(arr75), add=TRUE, col='blue')
```

---

## 2.1.4 Defining neighbours based on the optimisation of a trajectory

### About the travelling salesman

Some methods such as spatial sampling (see Chapter 10 "Spatial sampling") require prior data sorting. When the latter are characterised by two variables (*i.e.* their geographical coordinates in the plan), how to choose a sorting method becomes a complex theoretical problem.

One solution consists in running a *path* along all the points, and sorting them by their order of appearance when the path is taken. The neighbours of a given point are then the points located just before or just after along the path.

Out of the set of possible paths, some have characteristics that are better suited to the desired objectives, such as, for instance, reducing sampling variance. This is the case of the *shortest path*. It minimises the sum of the distances between two consecutive points. This path, which does not set any particular constraints on the starting or arrival point, is known in the literature of graph theory as the *Hamilton path* (Figure 2.11b) associated with a graph the edges of which are weighted.

A particular and well-known case of *shortest path* is that of the travelling salesman. It represents the path which a travelling salesman must take to visit all his customers, minimising the distance travelled and managing to return home in the evenings. Such a path corresponds to a Hamiltonian cycle (Figure 2.11c).

Looking for a shortest path is a classic optimisation problem in the context of graph theory. It

can be seen in particular in Euler's attempt to solve the problem of the seven Königsberg bridges<sup>1</sup>. It also plays a part in questions relating to Eulerian or Hamiltonian graphs<sup>2</sup>. Today, there are no algorithms in polynomial time that can be used to find the shortest path. When the number of points is high, the search for the optimal path requires the use of heuristics<sup>3</sup> resulting in a local optimum. They are available in package *TSP* in R (Hahsler et al. 2017).

When the distance is Euclidean and the number of points is reasonable, around a few hundreds, an exact solution can be found thanks to the *concorde* programme (Applegate et al. 2006). This programme can be called up directly from package *TSP* in R.

Lastly, the search for a Hamiltonian path from a distance matrix is equivalent to that of a Hamiltonian cycle, provided that a line and a column formed of 0 are added to the original matrix (Garfinkel 1985). Package *TSP* explicitly refers to this case with the *insert-dummy* function.

### Other methods

The *general randomized tessellation stratified* method (GRTS , Stevens Jr et al. 2004) is popular in spatial sampling, as it makes it possible to get a spatially-balanced sample for a finite population of individuals (distinct and identifiable units of dimension 0 of a discrete population, e.g. trees in a forest), a linear population (continuous units of dimension 1, e.g. rivers) or a population of surfaces (continuous units of dimension 2, e.g. forests). It is based on a path built from a class of functions referred to as *quadrant-recursive* (Mark 1990), making it possible to ensure that certain two-dimensional spatial proximity relationships are still preserved in one-dimensional space.

The idea of the method is to project the coordinates on a unit square, then cut this square into four cells, each of which is cut again into four sub-cells, etc. To each cell, a value is assigned, resulting from the order in which the division was carried out, ultimately making it possible for the units to be placed on the path going through the two-dimensional space.

Figure 2.10 shows the initial stages of cutting, which can be implemented with package *spsurvey* in R (Kincaid et al. 2016). However, with the GRTS method, *large jumps* (Figures 2.11d) are created along the paths, which can affect the accuracy of the estimates.

### Application with R - Source: Finding a shorter path

---

```
library(TSP)
library(miscTools)

#The utility software "concorde" must be downloaded at this address:

http://www.tsp.gatech.edu/concorde/downloads/downloads.htm
#and called from R

Sys.setenv(PATH=paste(Sys.getenv("PATH"), "z:/cygwin/App/Runtime/Cygwin/bin"
,sep=";"))
concorde_path("Z:/concorde/")

#The input data are a distance matrix
```

---

1. The issue studied by Euler was: in the city of Königsberg, is it possible to take a walk in which each of the 7 bridges is used once and only once? (**euler1741solutio**).

2. A Eulerian graph is a graph that can be travelled from a given vertex and walking along each edge exactly once before returning to the starting point vertex. It can be likened to a drawing that can be etched without ever lifting the pencil from the page. A Hamiltonian graph is a graph that can be travelled passing across all vertices and only once. A Hamiltonian graph is not necessarily Eulerian because in a Hamiltonian cycle, it is entirely possible to omit to pass through certain edges.

3. A heuristic is a calculation method that quickly (in polynomial time) provides a feasible solution, albeit not necessarily optimal.

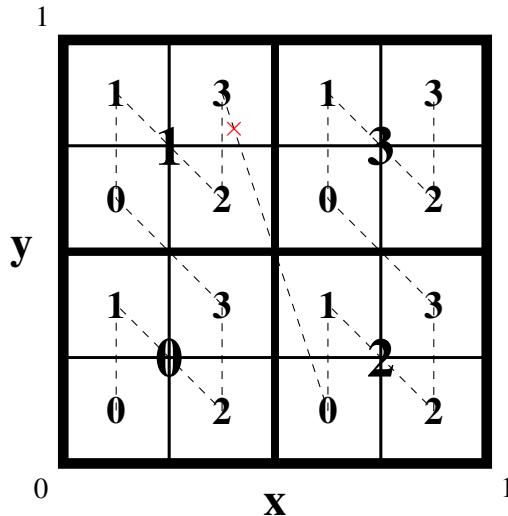


Figure 2.10 – Constructing a path with the GRTS method

**Note:** The value "13" is associated with the unit the position of which is a red cross, thus making it possible to position it on the path.

```
test <-as.matrix(read.csv("U:/paris.csv",header=FALSE,sep="\t"))

#rounding errors can lead to the matrix not being completely symmetrical.

tsp <-(symMatrix(test[upper.tri(test, TRUE)], nrow=nrow(test), byrow=TRUE))
#an object readable by TSP is created
tsp<-TSP(tsp)
#The concorde method is applied to this object.
tour<-solve_TSP(tsp, method = "concorde")
```

## 2.2 Attributing weights to neighbours

### 2.2.1 From a list of neighbours to a weight matrix

Once the neighbourhood graph has been defined and codified into a list of neighbours, the link between points  $i$  and  $j$  is transformed into the element  $w_{ij}$  of the weight matrix  $\mathbf{W}$ . The weight matrix  $\mathbf{W}$  is the "formal expression of spatial dependency between observations" (Anselin et al. 1988).

#### Defining the weight matrix

- Most commonly, the weight matrix is a binary contiguity matrix (see Figure 2.12):

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked in space} \\ 0 & \text{otherwise.} \end{cases} \quad (2.3)$$

- The weight matrices can also take into account the distance between the geographical zones, as relationships becoming smaller with distance: 1 if  $d < d_0$  - 0 otherwise,  $\frac{1}{d^\alpha}$ , or  $e^{-\alpha d}$  with  $\alpha$  an estimated or predetermined parameter. Using a maximum distance beyond which  $w_{ij} = 0$  makes it possible to limit the number of components with a value different from zero. As described in 2.1.2, when the size of the zones is heterogeneous, this method increases the risk of a considerable variability in the number of neighbours.

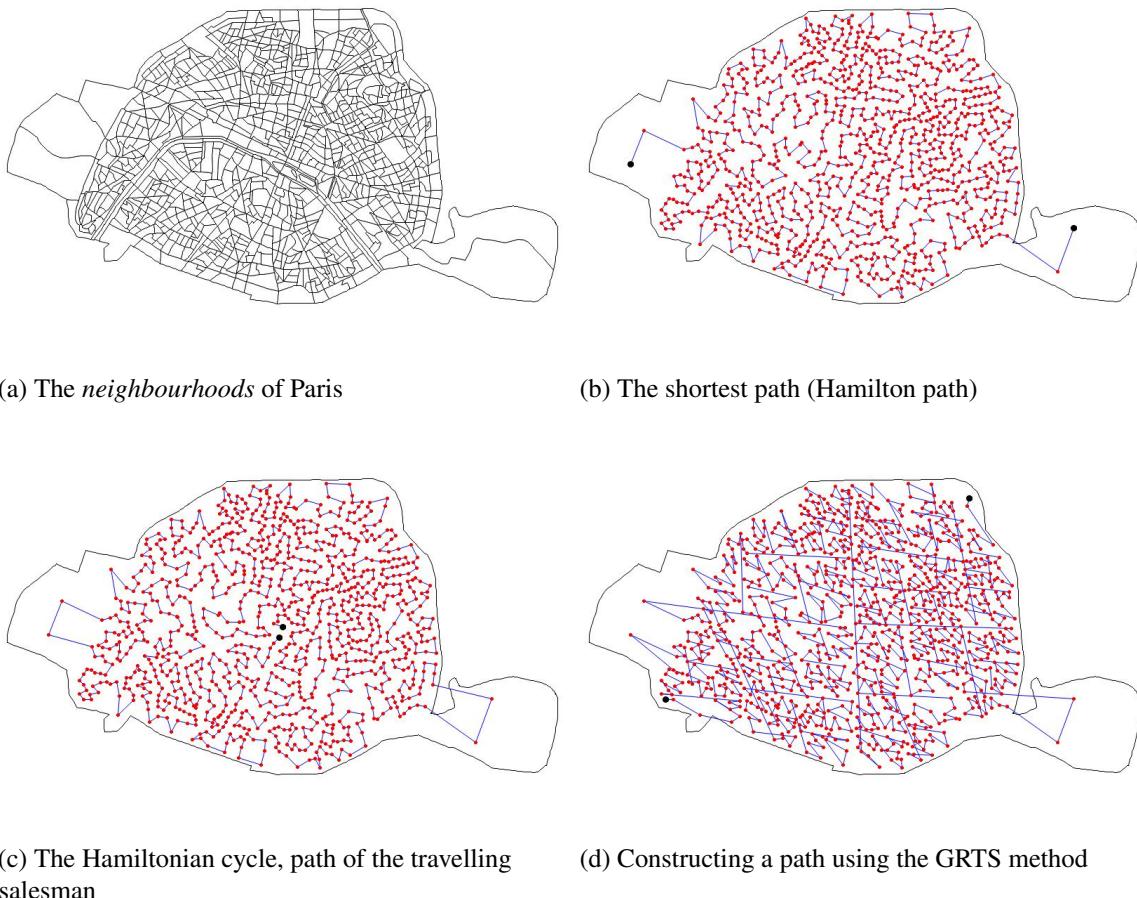


Figure 2.11 – Looking for paths that cross through all the *neighbourhoods* of Paris

					Sum of the neighbours weights						Sum of the neighbours weights							
<b>ROOK contiguity</b>					<b>QUEEN</b>													
	a	b	c	d	e		a	b	c	d	e		a	b	c	d	e	
a	0	1	1	0	0	2	0	1	1	1	0	3						
b	1	0	0	1	0	2	1	0	1	1	1	4						
c	1	0	0	1	0	2	1	1	0	1	0	3						
d	0	1	1	0	1	3	1	1	1	0	1	4						
e	0	0	0	1	0	1	0	1	0	1	0	2						

Figure 2.12 – Binary weight matrix

- Lastly, certain matrices take the strength of relations between the zones into account. For example, weight can be defined by  $\frac{b_{ij}^\alpha}{d_{ij}^\beta}$  with  $b_{ij}$  a measure of the strength of relationships between zones  $i$  and  $j$  (which is not necessarily symmetrical), such as the percentage of common boundaries, the total population, the wealth and  $d_{ij}$  the distance between the zones.

Some econometric studies are aimed at endogenising the weight matrices, but they are considered to be exogenous in most spatial econometric applications (Anselin 2013). In general, therefore, the neighbourhood weights must not be a function of the phenomenon which we are trying to explain.

### The "weight list" object in R

The function nb2listw of package spdep makes it possible to convert a "list of neighbours" object into a "weight list" object. It is important to note that the "weight list" object, which corresponds to the weight matrix described above, is not a matrix  $n \times n$  as represented in theory. It is a list containing the standardisation style and then for each observation: its attribute, the list of observation numbers of its neighbours, the list of the attributes of its neighbours and the list of the weights of its neighbours. Reference is often made to *sparse matrices*.

When a zone has no neighbours, the option zero.policy=TRUE makes it possible to generate a list of weights which takes value 'zero' for observations without neighbours (if the option is FALSE, an error message is generated).

### Application with R

---

```
#Matrix based on contiguity
#The function nb2listw converts any object of the nb type into a weight
list
arr75.lw <- nb2listw(arr75.nb)

#Matrix based on distance
#The mat2listw function converts a matrix into a weight list
library(fields) #to calculate the distance between two points
coords <- coordinates(arr75)
distance <- rdist(coords,coords)
diag(distance) <- 0
distance[distance >=100000] <- 0
#the weight decreases as a square of the distance, within a radius of 100
km
dist <- 1.e12 %/%
(distance*distance)
dist[dist >=1.e15] <- 0
dist.w <- mat2listw(dist, row.names=NULL)
```

---

### Weight matrix standardisation

The sum of the weights of the neighbours of a zone is called its *degree of connection*. If the weight matrix is not standardised ("B" coding scheme), the degree of connection will depend on the number of its neighbours, which creates heterogeneity between the zones. According to Tiefelsdorf 1998, four types of standardisation can be distinguished:

- Line standardisation ("W" coding scheme): for a given zone, the weight ascribed to each neighbour is divided by the sum of the weights of its neighbours:  $\sum_{j=1}^n w_{ij} = 1$ . This standardisation makes the interpretation of the weight matrix easier, because  $\sum_{j=1}^n w_{ij}x_j$  represents the average of variable  $x$  on all neighbours of observation  $i$ . Each weight  $w_{ij}$

can be interpreted as the fraction of spatial influence on observation  $i$  ascribable to  $j$ . In contrast, such standardisation implies a certain degree of competition between neighbours: the fewer neighbours a zone has, the greater their weight. Moreover, when weights are inversely proportional to the distance between the zones, row standardisation makes them difficult to interpret.

— Global standardisation ("C" coding scheme): weights are standardised so that the sum of all weights is equal to the total number of entities. All weights are multiplied by  $\frac{n}{\sum_{j=1}^n \sum_{i=1}^n w_{ij}}$ .

— Uniform standardisation ("U" coding scheme): weights are standardised so that the sum of all weights equals 1:  $\sum_{j=1}^n \sum_{i=1}^n w_{ij} = 1$ .

— Standardisation by variance stabilisation ("S" coding scheme): let  $\mathbf{q}$  be the vector defined by:

$$\mathbf{q} = (\sqrt{\sum_{j=1}^n w_{1j}^2}, \sqrt{\sum_{j=1}^n w_{2j}^2}, \dots, \sqrt{\sum_{j=1}^n w_{nj}^2})^T.$$

Let matrix  $\mathbf{S}^* = [\text{diag}(\mathbf{q})]^{-1} \mathbf{W}$ .<sup>4</sup> From  $\mathbf{S}^*$ , we calculate  $Q = \sum_{j=1}^n \sum_{i=1}^n s_{ij}^*$  from which we deduce the standardised weight matrix:  $\mathbf{S} = \frac{n}{Q} \mathbf{S}^*$ .

Standardisation by variance stabilisation was introduced by Tiefelsdorf in order to reduce the heterogeneity in the weights due to differences in size and the number of neighbours between zones. Line standardisation gives more weight to observations bordering the study zone, with a small number of neighbours. On the contrary, with global or uniform standardisation, the observations in the centre of the study zone, with a large number of neighbours, are subject to more external influences than the border zones. This heterogeneity can have a significant impact on the results of spatial autocorrelation tests.

The weight of the standardised matrix based on the "S" coding scheme varies less than those of the standardised matrix based on the "W" scheme. The sum of the weights of the lines varies more for the "S" scheme than for the "W" scheme, but less than for the "B", "C" and "U" schemes (Bivand et al. 2013b).

Whether the coding scheme is in row, global, or by variance stabilization, the sum of all elements in the matrix is always  $n$ , which enables the spatial autocorrelation statistics using the matrix to be comparable to each other.

### Application with R

---

```
#The style option makes it possible to set the type of standardisation
arr75.lw <- nb2listw(arr75.nb,zero.policy=TRUE, style="W")
names(arr75.lw)
## [1] "style"      "neighbours" "weights"
summary(unlist(arr75.lw$weights))
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
## 0.1250 0.1667 0.1833 0.1961 0.2500 0.3333
```

---

## 2.2.2 Importance of the choice of weight matrix

When trying to test the importance of economic or social relationships between certain variables, the geographical location of the observations is a key parameter. First of all, observations in the same geographical zone are subject to the same external parameters (climate, pollution, etc.) Secondly, neighbouring observations mutually influence one another. Spatial econometrics models take these various interactions into account. These models use neighbourhood specification *via* weight matrix

4.  $\text{diag}(\mathbf{q})$  is a diagonal matrix with the components of  $\mathbf{q}$  on its main diagonal

W. Within the scientific community, opinions diverge on the influence of the definition of the weight matrix on results.

Bhattacharjee et al. 2005 note that: "The choice of weights is often arbitrary [...] and the result of the studies varies considerably depending on the definition of the spatial weights". A poor specification of  $\mathbf{W}$  would lead to false conclusions. Having said that, as different weight matrix construction methods can be applied, "[...] it is possible that one method leads to relevant results, though the risk of a poor specification will always weigh on the chosen model". (Getis et al. 2004).

The aim is that the weights  $w_{ij}$  reflect interactions between observations as accurately as possible. The underlying assumptions can be based on economic or sociological models. For example, zero weight beyond a certain distance will be justified by the fact that the influence of an employment area on its environment is constrained by the mobility of individuals, which is itself limited by their travelling time. However, Harris et al. 2011 emphasise that the concept of 'distance' is itself unclear. Distance is often defined by a geometric distance between two representative points of the study zones. But distance can also be the transport time between two regions (minimum time, or time taking the least expensive route), or for instance be proportional to interactions between zones. According to Harris et al. 2011, "the consequence of using measures connected with contiguity or distance to weight the observations of neighbouring regions is that a spatial interaction structure is imposed without any means of verifying its reliability, such that it may be poorly specified."

Harris et al. 2011 show some alternative approaches to weight matrix construction. These methods aim at minimising the *ad hoc* hypotheses in matrix specification. However, no method gets rid of it completely.

Not all researchers are as pessimistic: LeSage et al. 2010 consider that the belief that weight matrix has a crucial influence on results is due to errors in interpreting the coefficients of spatial econometrics models, or to errors in model specification. In their words, this belief is "the biggest myth in spatial econometrics". They argue that if we look at the average effect of explanatory variables on dependent variables, the differences in weight matrix specification do not have a significant influence on results. However, Lesage et al. 2009 acknowledge that much remains to be done toward better characterising the concept of equivalence between matrices.

## References - Chapter 2

- Anselin, Luc (2013). *Spatial econometrics: methods and models*. Vol. 4. Springer Science & Business Media.
- Anselin, Luc and Daniel A Griffith (1988). « Do spatial effects really matter in regression analysis? ». *Papers in Regional Science* 65.1, pp. 11–34.
- Applegate, David et al. (2006). *Concorde TSP solver*.
- Bhattacharjee, Arnab and Chris Jensen-Butler (2005). « Estimation of spatial weights matrix in a spatial error model, with an application to diffusion in housing demand ». CRIEFF Discussion Papers.
- Bivand, Roger S, Edzer Pebesma, and Virgilio Gomez-Rubio (2013b). « Spatial Neighbors ». *Applied Spatial Data Analysis with R*. Springer, pp. 83–125.
- Garfinkel, R.S. (1985). « Motivation and modelling (chapter 2) ». *E. L. Lawler, J. K. Lenstra, A.H.G. Rinnooy Kan, D. B. Shmoys (eds.) The traveling salesman problem - A guided tour of combinatorial optimization*, Wiley & Sons.
- Getis, A and J Aldstadt (2004). « On the specification of the spatial weights matrix ». *Geographical Analysis* 35.
- Hahsler, Michael and Kurt Hornik (2017). *TSP: Traveling Salesperson Problem (TSP)*. R package version 1.1-5. URL: <https://CRAN.R-project.org/package=TSP>.
- Harris, Richard, John Moffat, and Victoria Kravtsova (2011). « In search of 'W' ». *Spatial Economic Analysis* 6.3, pp. 249–270.
- Kincaid, Thomas M. and Anthony R. Olsen (2016). *spsurvey: Spatial Survey Design and Analysis*. R package version 3.3.
- LeSage, James P and R Kelley Pace (2010). « The biggest myth in spatial econometrics ». Available at SSRN 1725503.
- Lesage, James and Robert K Pace (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.
- Mark, David M (1990). « Neighbor-based properties of some orderings of two-dimensional space ». *Geographical Analysis* 22.2, pp. 145–157.
- Smith, Tony E. (2016). *Notebook on Spatial Data Analysis*. <http://www.seas.upenn.edu/ese502/notebook>.
- Stevens Jr, Don L and Anthony R Olsen (2004). « Spatially balanced sampling of natural resources ». *Journal of the American Statistical Association* 99.465, pp. 262–278.
- Tiefelsdorf, Michael (1998). « Modelling spatial processes: The identification and analysis of spatial relationships in regression residuals by means of Moran's I (Germany) ». PhD thesis. Université Wilfrid Laurier.
- Toussaint, Godfried T (1980). « The relative neighbourhood graph of a finite planar set ». *Pattern recognition* 12.4, pp. 261–268.
- (2014). « The sphere of influence graph: Theory and applications ». *International Journal of Information Technology and Computer Science* 14.2.





## Part 2: Measuring the importance of spatial effects

3	Spatial autocorrelation indices .....	51
4	Spatial distribution of points .....	71
5	Geostatistics .....	113



### 3. Spatial autocorrelation indices

BOUAYAD AGHA SALIMA

GAINS (TEPP) et CREST

*Le Mans Université*

MARIE-PIERRE DE BELLEFON

INSEE

---

<b>3.1</b>	<b>What is spatial autocorrelation?</b>	<b>52</b>
3.1.1	Empirical observation of spatial autocorrelation .....	52
3.1.2	Moran's diagram .....	53
<b>3.2</b>	<b>Global measures of spatial autocorrelation</b>	<b>55</b>
3.2.1	Spatial autocorrelation indices .....	55
3.2.2	Spatial autocorrelation of categorical variables .....	60
<b>3.3</b>	<b>Local measures of spatial autocorrelation</b>	<b>62</b>
3.3.1	Getis and Ord index .....	62
3.3.2	Local spatial autocorrelation indicators .....	63
3.3.3	Significance of the local Moran's I .....	63
3.3.4	Interpretation of local indices .....	67
<b>3.4</b>	<b>Spatio-temporal indices</b>	<b>68</b>

---

#### Abstract

Spatial autocorrelation indices measure the spatial dependence between values of the same variable in different places in space. The more the observation values are influenced by observation values that are geographically close to them, the greater the spatial correlation.

This chapter defines spatial autocorrelation, then describes spatial autocorrelation indices at the global and local levels — principles, properties, practical implementation with R and interpretation of their significance.

- R** Prior reading of Chapters 1 "Descriptive spatial analysis" and 2 "Codifying the neighbourhood structure" is recommended.

Very often, the variables for which geolocated information is available are characterised by spatial dependencies, which are all the stronger as the locations are closer. Thus, increasingly frequent access to spatial data makes it possible to better take into account interactions and spatial externalities in analysing the economic decisions made by agents. Analysis of spatial structures included in the data is essential for addressing, if necessary, any violation of the hypothesis of spatial independence of variables. Secondly, when it comes to interpretation, the analysis of spatial autocorrelation enables quantified analysis of the spatial structure of the phenomenon in question. Spatial autocorrelation indices measure the spatial dependence between values of the same variable in different places in space.

### 3.1 What is spatial autocorrelation?

Autocorrelation measures the correlation of a variable with itself, when the observations are considered with a time lag (temporal autocorrelation) or in space (spatial autocorrelation). Spatial autocorrelation is defined as the positive or negative correlation of a variable with itself due to the spatial location of the observations. This spatial autocorrelation can first be the result of unobservable or difficult-to-quantify processes that combine different locations and, as a result, give rise to a spatial structuring of activities: interaction phenomena – between agents' decisions, for example – or dissemination – such as phenomena of technological diffusion – in space are each phenomena that can produce spatial autocorrelation. Secondly, in the context of the specification of econometric models, measuring spatial autocorrelation can be considered a tool for diagnosing and detecting an incorrect specification – omission of spatial variables that are spatially correlated, errors on the choice of scale on which the spatial phenomenon is analysed, etc.

From a statistical point of view, many analyses – analysis of correlations, linear regressions, etc. – are based on the hypothesis of independence of variables. When a variable is spatially auto-correlated, the independence hypothesis is no longer respected, thus challenging the validity of the hypotheses on the basis of which these analyses are carried out. Secondly, analysis of spatial autocorrelation enables quantified analysis of the spatial structure of the studied phenomenon.

It should be emphasised that spatial structure and spatial autocorrelation cannot exist independently of one another (Tiefelsdorf 1998):

- The term spatial structure refers to all the links with which the autocorrelated phenomenon will spread;
- without the presence of a significant autocorrelated process, the spatial structure cannot be empirically observed.

The spatial distribution observed is then considered the manifestation of the underlying spatial process.

#### 3.1.1 Empirical observation of spatial autocorrelation

In the presence of spatial autocorrelation, it is observed that the value of a variable for an observation is linked to the values of the same variable for the neighbouring observations.

- Spatial autocorrelation is positive when similar values of the variable to be studied are grouped geographically.
- Spatial autocorrelation is negative when the dissimilar values of the variable to be studied come together geographically — nearby locations are more different than remote locations. This type of situation is usually found in the presence of spatial competition.

- In the absence of spatial autocorrelation, it can be considered that the spatial allocation of the observations is random.

Spatial autocorrelation indices make it possible to assess spatial dependence between values of the same variable in different places in space and test the significance of the identified spatial structure. To show this, the indices take into account two criteria:

- spatial proximity;
- the similarity or dissimilarity of the values of the variable for the spatial units considered.

Beware: if the data are aggregated following a breakdown that does not respect the underlying phenomenon, the strength of the spatial link will be overestimated or underestimated.

The measurement of a global spatial autocorrelation is distinguished from that in a given space and local autocorrelation in each unit of this space. This measures the intensity and significance of local dependence between the value of a variable in a spatial unit and the values of the same variable in neighbouring units (more or less close, depending on the neighbourhood criterion used).

### 3.1.2 Moran's diagram

Moran's diagram allows a rapid reading of the spatial structure. This is a scatter graph with the values of variable  $y$  centred on the x-axis and the average values of the variable for the neighbouring observations  $Wy$  in the y-axis, where  $W$  is the normalized weight matrix. The two properties, *y centred* and *W normalized* imply that empirical average  $Wy$  is equal to that of  $y$  and therefore 0. The straight regression line of  $Wy$  is also drawn depending on  $y$  and equation lines  $y = 0$  and  $Wy = 0$  that delineate the quadrants.

If the observations are randomly distributed in space, there is no particular relationship between  $y$  and  $Wy$ . The slope of the linear regression line is zero, and the observations are evenly allocated in each quadrant. If, on the contrary, observations have a particular spatial structure, the linear regression slope is non-null since there is a correlation between  $y$  and  $Wy$ . Each of the quadrants defined by  $y = 0$  and  $Wy = 0$  matches up with a type of specific space association (Figures 3.1 and 3.2).

- The observations in the upper right – quadrant 1 – show values of the variable that are higher than average, in a neighbourhood similar to it — positive spatial autocorrelation and high index value; high-high structure.
- At the bottom left – quadrant 3 – the observations show lower variable values than average, in a neighbourhood similar to it — positive space autocorrelation and low index value; low-low structure.
- Observations located at the bottom right – quadrant 2 – have higher values of the variable than average in a neighbourhood not similar to it — negative spatial autocorrelation and high index value; high-low structure.
- At the top left – quadrant 4 – the observations show values for the variable that are lower than the average in a neighbourhood not similar to it — negative spatial autocorrelation and low index value; low-high structure.

The density of points in each of the quadrants is used to visualise the dominant spatial structure. Moran's diagram also makes it possible to see the atypical points that move away from this spatial structure.

To understand how spatial autocorrelation can be seen on Moran's diagram, we simulated a growing spatial autocorrelation of incomes by IRIS (Figures 3.3 and 3.4). Parameter  $\rho$  that defines spatial autocorrelation is the slope in Moran's chart. Apart from extreme values, it is difficult to identify the sign and the strength of spatial autocorrelation by simply looking at the maps of the various values. On the other hand, Moran's diagrams make it possible to clearly identify the various scenarios.

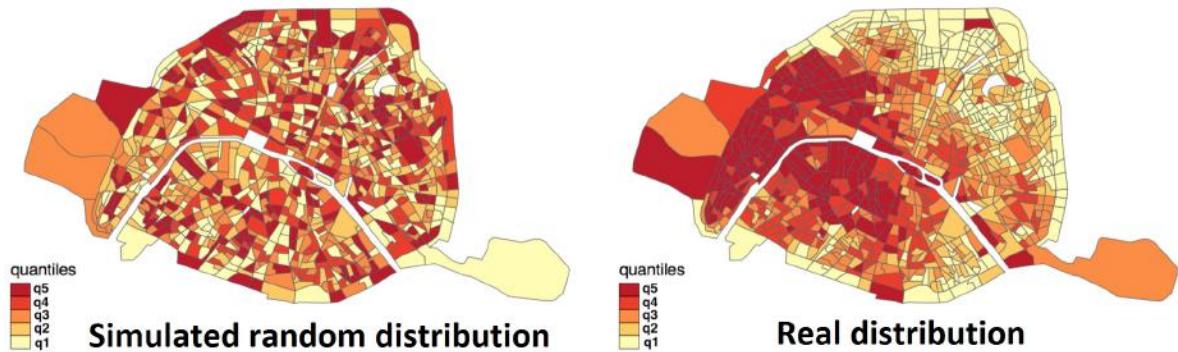


Figure 3.1 – Illustration, on Parisian census districts (IRIS), of the gap between random distribution and spatially autocorrelated distribution

Source: INSEE, Localised Tax Revenues System (RFL) 2010

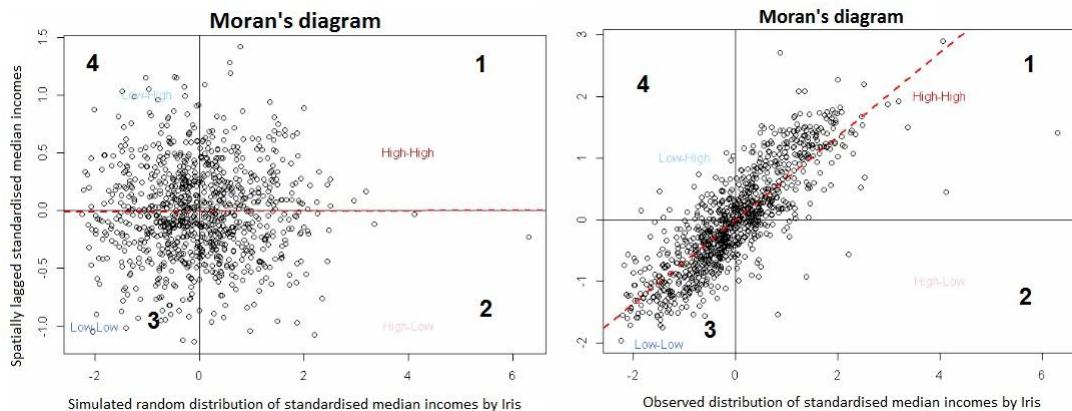


Figure 3.2 – Moran's diagram of a simulated random distribution of standardised median incomes by IRIS and standardised median incomes by IRIS

Source: INSEE, Localised Tax Revenues System (RFL) 2010

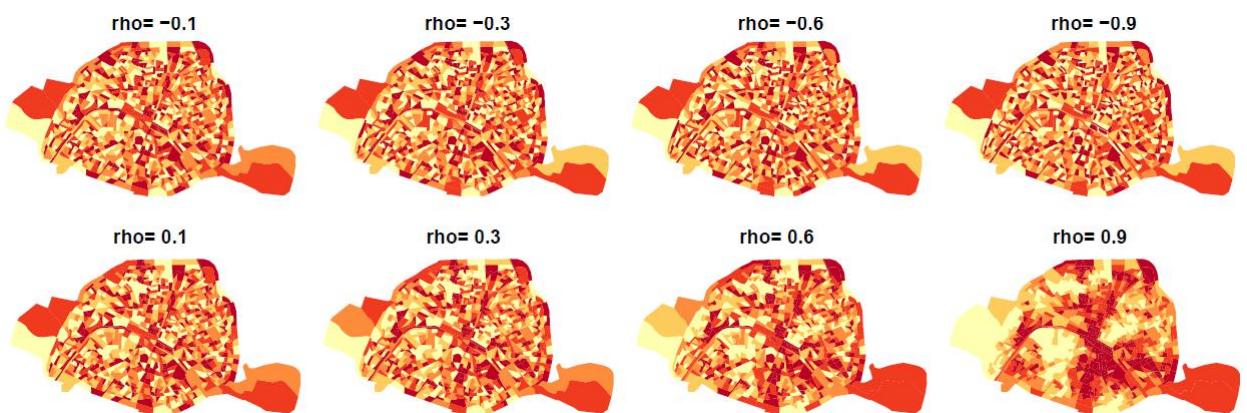


Figure 3.3 – Simulation of increasing spatial autocorrelation of incomes by IRIS

Source: INSEE, Localised Tax Revenues System (RFL) 2010

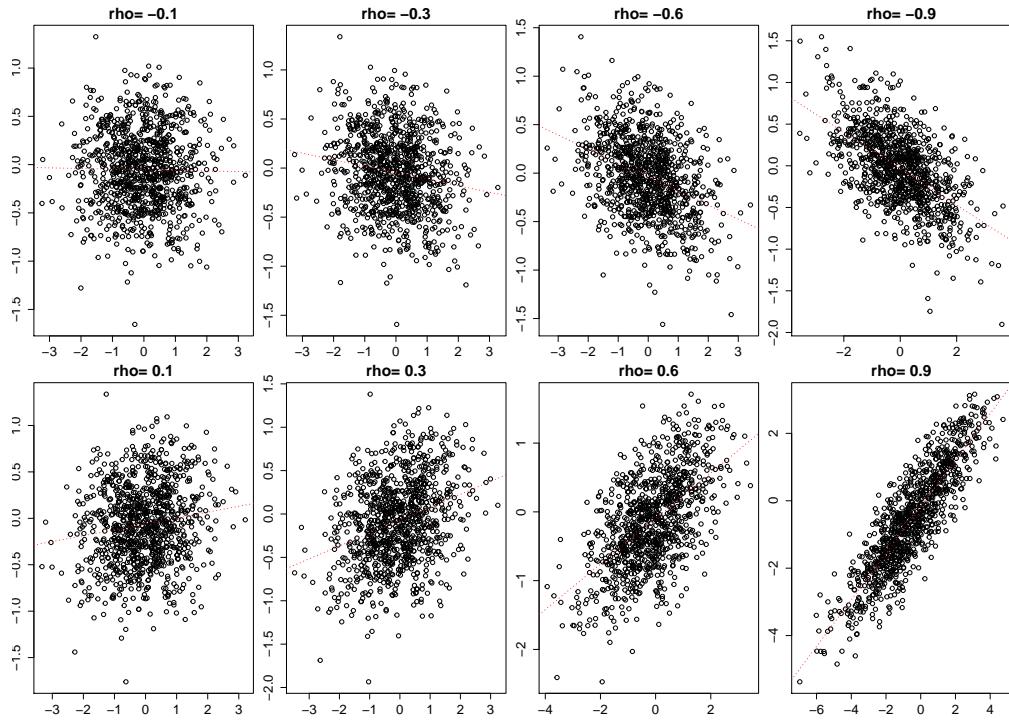


Figure 3.4 – Moran diagrams according to simulated autocorrelated income, for Parisian IRIS  
Source: INSEE, Localised Tax Revenues System (RFL) 2010

## 3.2 Global measures of spatial autocorrelation

### 3.2.1 Spatial autocorrelation indices

When Moran's diagram highlights a particular spatial structure, calculating the spatial autocorrelation indices is used to answer two questions:

- Could the values taken by the neighbouring observations have been comparable (or also dissimilar) by mere chance?
- If not, then we are dealing with a case of spatial autocorrelation. How is this denoted and what is the strength of the said autocorrelation?

To answer the first question, we must test the hypothesis of absence of spatial autocorrelation for a gross variable  $y$ .

- $H_0$ : no spatial autocorrelation
- $H_1$ : spatial autocorrelation

To carry out this test, it is necessary to specify the distribution of the variable of interest  $y$ , in the absence of spatial autocorrelation (under  $H_0$ ). In this context, statistical inference is generally conducted considering either of the following assumptions:

**Normality hypothesis:** each of the values of the variable, or  $y_i$ , is the result of an independent draw in the **normal distribution specific to each geographical area  $i$  on which this variable is measured**.

**Randomisation hypothesis:** The inference over Moran's I is usually conducted under the randomisation hypothesis. The estimated statistic calculated from data is compared with the **distribution of the data derived by randomly re-ordering the data – permutations**. The idea is simply that if the null hypothesis is true, then all possible combinations of data are equiprobable. The data observed are then only one of the many outcomes possible. In the case of spatial autocorrelation, the null hypothesis is always that there is no spatial association and the values of the variable are

randomly assigned to the spatial units in order to calculate the test statistic. If the null hypothesis is rejected, *i.e.* if spatial autocorrelation is found, we can then calculate the range of values that governs the spatial autocorrelation index and thus answer the question as to the signals and strength of the spatial autocorrelation: the closer this index is to 1 in absolute value, the greater is the correlation. This interval depends on the weight matrix and can sometimes vary outside the interval  $[-1; 1]$ , hence the importance of calculating the limits of this interval.

Very generally speaking, spatial autocorrelation indices are used to characterise the correlation between measures that are geographically similar to a measured phenomenon. If  $WY$  is the vector of means of variable  $Y$  (where  $W$  is the spatial weights matrix) in the neighbourhood of each spatial unit, spatial autocorrelation indices occur as:

$$\text{Corr}(Y, WY) = \frac{\text{Cov}(Y, WY)}{\sqrt{\text{Var}(Y) \cdot \text{Var}(WY)}} \quad (3.1)$$

Based on this very general formulation, for quantitative variables, two main indices are used to test for spatial autocorrelation — the Moran index and the Geary index. The former considers the variances and covariances taking into account the difference between each observation and the average of all observations. The Geary index takes into account the difference between the respective observations. In the literature, Moran's index is often preferred to that of Geary due to greater general stability (see in particular Upton et al. 1985).

### Moran index

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \quad (3.2)$$

- $H_0$ : The neighbours do not **co-vary** in any particular way.
- $I_W > 0 \Rightarrow$  positive spatial autocorrelation.

### Geary index

$$c_W = \frac{n-1}{2 \sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - y_j)^2}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \quad (3.3)$$

- $H_0$ : The **differences** between neighbours have no particular structure.
- $c_W < 1 \Rightarrow$  positive spatial autocorrelation.

Depending on the distribution chosen for the variable in the absence of spatial autocorrelation, the calculation of the variance of the indices is modified. In contrast, the equations that yield the expression of the expectancy of the indices (3.4) and the test statistic (3.5) remain the same. These relationships thus make it possible to assess the significance of spatial autocorrelation.

$$E(I_W) = E(c_W) = -\frac{1}{n-1} \quad (3.4)$$

$$\frac{I_W - E(I_W)}{\sqrt{\text{Var}(I_W)}} \sim \frac{c_W - E(c_W)}{\sqrt{\text{Var}(c_W)}} \sim \mathcal{N}(0, 1) \quad (3.5)$$

As spatial autocorrelation is measured based on a comparison of the value of an individual variable with that of its neighbours, the definition of the neighbourhood will have a significant impact on the measurement of spatial autocorrelation. As explained in Chapter 2 "Codifying the neighbourhood structure", the larger the planned neighbourhood, the greater the number of neighbours considered, and the greater the probability that their average will be closer to the population's average, which may lead to a relatively low value for spatial autocorrelation.

A change in scale can also have implications when measuring spatial autocorrelation. The term MAUP (Modifiable Areal Unit Problem; Openshaw et al. 1979b) is used to describe the influence of spatial breakdown on the results of statistical processing or modelling.

More precisely, the irregular forms and limits of the administrative levels that do not necessarily reflect the reality of the spatial distributions studied are an obstacle to the comparability of the irregularly distributed spatial units. According to Openshaw 1984, MAUP is a combination of two distinct but similar problems:

- The scale problem stems from a change in the information generated when a set of spatial units is aggregated to form smaller and larger units for the needs of an analysis or due to data availability issues;
- The aggregation problem – or zoning – stems from a change in the diversity of information generated by the various aggregation schemes possible at a same scale. This effect is characteristic of administrative partitioning – particularly electoral – and adds to the scale effect.

■ **Example 3.1 — Spatial autocorrelation of median income in Paris.** What is the intensity of spatial autocorrelation in the income of Parisians? Is it significant? To what extend does it depend on the specification of spatial relations – type of neighbourhood, aggregation scale –?

### Measuring spatial autocorrelation

Source	$I_W$	$c_w$	p value	H0	limits of $I_W$
Income: breakdown observed	0.68	0.281	$3.10^{-6}$	rejected	[-1.06,1.06]
Income: simulated random distribution	0.0027	1.0056	> 0.5	accepted	[-1.06,1.06]

Table 3.1 – Moran and Geary indices of the median income earned by Parisians by IRIS:  
observed and simulated distribution

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

### Influence of neighbourhood choice

Type of neighbourhood	$I_W$	p value	H0
QUEEN	0.68	$3.10^{-6}$	rejected
ROOK	0.57	$2.10^{-6}$	rejected
1NN	0.30	0.07	rejected
3NN	0.58	$9.10^{-6}$	rejected
Delaunay	0.57	$6.10^{-7}$	rejected

Table 3.2 – Moran and Geary indices of the median income earned by Parisians by IRIS  
according to the neighbourhood defined

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

### Influence of the aggregation scale

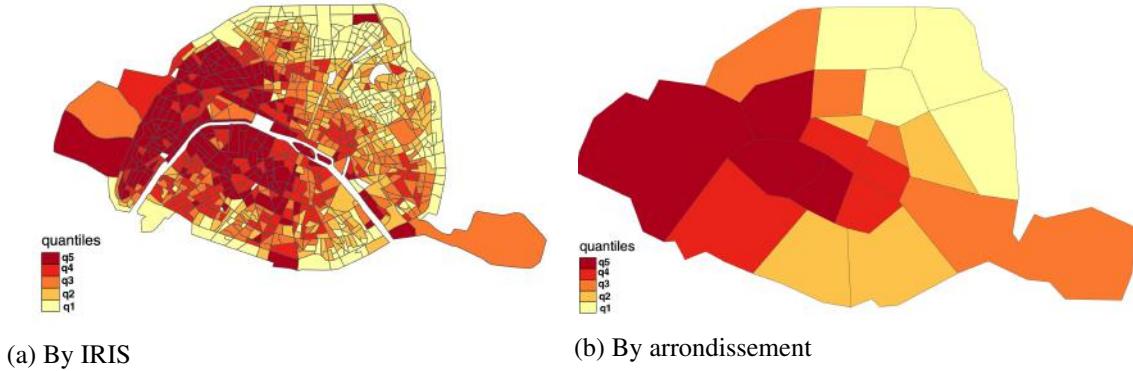


Figure 3.5 – Aggregation of income in Paris

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

Aggregation scale	$I_W$	p value	H0	Boundaries of $I_W$
IRIS	0.68	$3.10^{-6}$	rejected	[-1.06,1.06]
Arrondissement	0.51	$<9.10^{-9}$	rejected	[-0.53,1.01]

Table 3.3 – Value and significance of the Moran's I as a function of the chosen aggregation scale

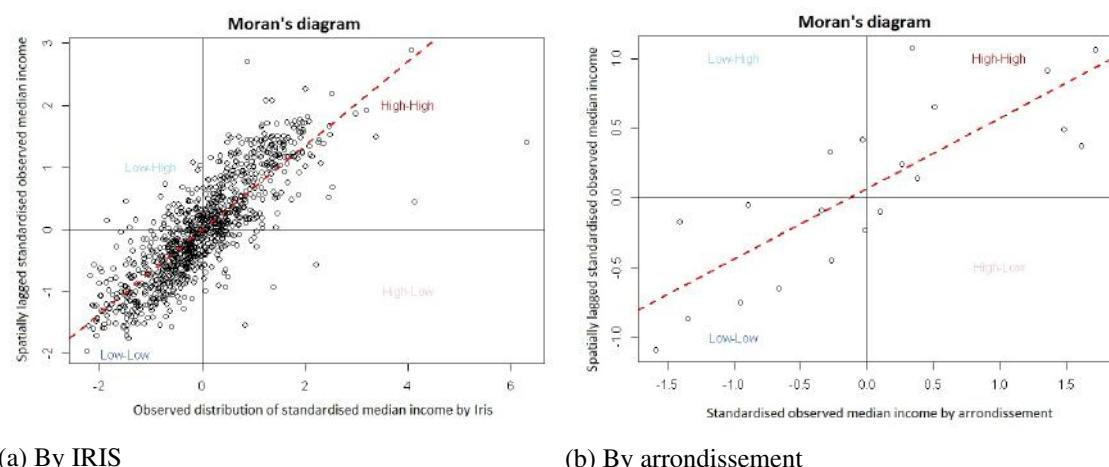
**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

Figure 3.6 – Moran's scatterplots for income distribution in Paris

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

In this example, whatever the definition of the neighbourhood or the aggregation scale, the spatial autocorrelation of the income earned by Parisians is positive and significant. The strength of spatial autocorrelation varies slightly depending on the type of neighbourhood used. In particular, looking only at the nearest neighbours slightly decreases the strength of spatial autocorrelation in this example.

■

### Application with R

The *spdep* package is used to calculate spatial autocorrelation indices and their significance using functions `moran.test` and `geary.test`.

By default, the distribution of the variable of interest under the null hypothesis is derived by randomisation. The `randomisation = FALSE` argument makes it possible to assume that this is a normal distribution.

**Box 3.2.1 — If certain entities do not have neighbours.** In order for the package functions *spdep* to accept spatial weight matrices in which certain units do not have neighbours, it is necessary to specify the option: `zero.policy=TRUE`. By default, the size of the matrix is reduced to exclude observations without neighbours. The opposite can be specified with the option: `adjust.n=FALSE`. In this case, the absolute value of the test statistic increases, and the absolute value of its expected maturity and variance decreases (Bivand et al. 2013a). Generally speaking, spatial autocorrelation indices were developed assuming that all units had neighbours, and there are different opinions on what to do when this is not the case.

As seen before, two approaches are used to estimate the significance of these indices — an analytical solution based on the normality hypothesis and a Monte Carlo solution based on the randomisation hypothesis. The analytical solution, used by the `moran.test` function, is based on the assumption that the test statistic asymptotically follows a normal distribution with mean 0 and variance 1. This is not always the most accurate measure of significance as convergence towards this distribution may depend on the arrangement of the polygons. Instead, the `moran.mc` function can be used, allowing to choose the number of permutations to calculate the simulated distribution of Moran's I. Comparing the significance levels calculated from functions `moran.mc` and `moran.test` makes it possible to ensure the robustness of the conclusions.

---

```
library(spdep)

#####
# Data preparation #####
#####

#Extraction of list of neighbours (defined by default with Queen contiguity
)
iris75.nb <- poly2nb(iris75)
#Creation of weight matrix
iris75.lw <- nb2listw(iris75.nb,zero.policy=TRUE)
#Calculation of standardised median income
iris75.data <- as.data.frame(iris75)
iris75.data$med_revenu_std <- scale(iris75.data$med_revenu)

#####
# Moran's diagram #####
#####
```

```

moran.plot(iris75.data$med_revenu_std, iris75.lw, labels=FALSE,
xlab='observed distribution of standardised median income by IRIS', ylab=
  'Spatially lagged standardised median incomes')

#####
# Moran's I test
#####

moran.test(iris75.data$med_revenu_std,iris75.lw, zero.policy=TRUE,
  randomisation=FALSE)

#Calculation of the range of Moran's I
moran.range <- function(lw) {
  wmat <- listw2mat(lw)
  return(range(eigen((wmat+t(wmat))/2)$values)))
}

moran.range(iris75.lw)

```

---

### 3.2.2 Spatial autocorrelation of categorical variables

When the variable of interest is not continuous but categorical, the degree of local association is measured by analysing the statistics of the *join count* (Zhukov 2010).

To illustrate the calculation of these statistics, we consider a binary variable representing two colours, White (B) and Black (N) so that a relation can be called White-White, Black-Black or White-Black. It can be seen that:

- positive spatial autocorrelation occurs if the number of White-Black relations is significantly **lower than** what would have occurred with random spatial distribution;
- negative spatial autocorrelation occurs if the number of White-Black relations is significant **greater than** what would have occurred with random spatial distribution;
- no positive spatial autocorrelation occurs if the number of White-Black links is approximately **identical to** what would have occurred with random spatial distribution;

If there are  $n$  observations,  $n_b$  white observations and  $n_n = n - n_b$  black observations, the probability of a white observation occurring is:  $P_b = \frac{n_b}{n}$  and the likelihood of a black observation occurring is:  $P_n = 1 - P_b$ .

In the absence of spatial autocorrelation, the probabilities of observations of the same colour occurring in two neighbouring cells are:  $P_{bb} = P_b * P_b = P_b^2$  and  $P_{nn} = P_n * P_n = (1 - P_b)^2$ .

The probability of obtaining different colour observation occurring in two neighbouring cells is:  $P_{bn} = P_b * (1 - P_b) + (1 - P_b) * P_b = 2P_b * (1 - P_b)$ .

As  $\frac{1}{2} \sum_i \sum_j w_{ij}$  measures the number of existing relations, assuming random spatial distribution

of the observations, it can be asserted that:

$$\begin{aligned} E[bb] &= \frac{1}{2} \sum_i \sum_j w_{ij} P_b^2 \\ E[nn] &= \frac{1}{2} \sum_i \sum_j w_{ij} (1 - P_b)^2 \\ E[bn] &= \frac{1}{2} \sum_i \sum_j w_{ij} 2P_b * (1 - P_b) \end{aligned} \quad (3.6)$$

Assuming  $y_i = 1$  when the observation is black and  $y_i = 0$  in the opposite case (white colour), the empirical counterparts (observed values) of these mathematical expectations can be written:

$$\begin{aligned} nn &= \frac{1}{2} \sum_i \sum_j w_{ij} y_i y_j \\ bb &= \frac{1}{2} \sum_i \sum_j w_{ij} (1 - y_i)(1 - y_j) \\ bn &= \frac{1}{2} \sum_i \sum_j w_{ij} (y_i - y_j)^2 \end{aligned} \quad (3.7)$$

In this case, the test statistic to assess the significance of spatial autocorrelation is based on the assumption that in the absence of spatial autocorrelation, the statistics of *joincount* (*bb*, *nn* and *bn*) follow a normal distribution. It can be written that:

$$\frac{bn - E(bn)}{\sqrt{Var(bn)}} \sim \mathcal{N}(0, 1) \quad \frac{bb - E(bb)}{\sqrt{Var(bb)}} \sim \mathcal{N}(0, 1) \quad \frac{nn - E(nn)}{\sqrt{Var(nn)}} \sim \mathcal{N}(0, 1) \quad (3.8)$$

■ **Example 3.2 — Joincount statistics on the employment of individuals in Paris.**<sup>1</sup>

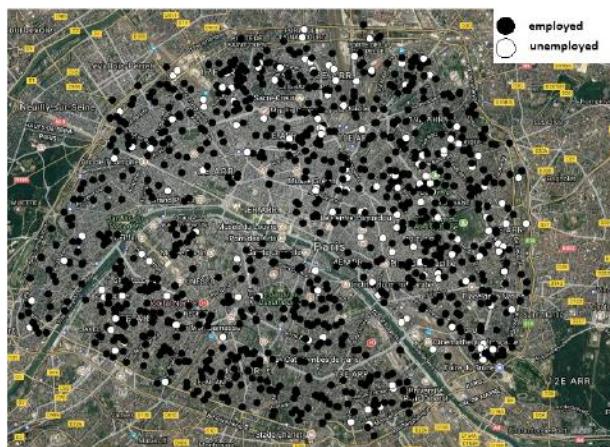


Figure 3.7 – Employment of a sample of 1,000 individuals in Paris

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

The binary variable is considered to be 1 if individual  $i$  is unemployed and 0 otherwise. The aim is to determine whether unemployed Parisians are more grouped into space than if they were randomly distributed. *Join count* statistics make it possible to answer this question. From the table 3.4, the location of the unemployed can be observed to be significantly correlated, as is the location of the employed.

1. The purpose of this example is not to detail the results of an economic study, but to illustrate the techniques implemented. There is no interpretation to be derived from this.

Variable	p-value of the join count statistic of spatial association	H0
Unemployed	$5.439 \cdot 10^{-3}$	rejected
Active workers	$9.085 \cdot 10^{-5}$	rejected

Table 3.4 – Significance of the join count statistic of Parisian unemployed

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

### Application with R

*Join count* statistic is reached by implementing `joincount.test` function of the `spdep` package in R.

---

```
library(spdep)

# Conversion as factor
menir10_d75_subset$unemployment <- ifelse(menir10_d75_subset$ZCHOM>0, 3, 1)
unemployment <- as.factor(menir10_d75_subset$unemployment, levels=c("employed", "unemployed"))

# Neighbours list and spatial weight matrices
coordinates(menir10_d75_subset) <- c("PLG_X", "PLG_Y")
proj4string(menir10_d75_subset) <- CRS("+init=epsg:27572 +proj=lcc +lat_1=46.8 +lat_0=46.8 +lon_0=0 +k_0=0.99987742 +x_0=600000 +y_0=2200000 +a=6378249.2 +b=6356515 +towgs84=-168,-60,320,0,0,0,0 +pm=paris +units=m +no_defs")
menir10_d75_subset <- spTransform (menir10_d75_subset, CRS ("+init=epsg :2154"))

menir75.nb <- knn2nb(knearneigh(menir10_d75_subset,k=2))

# Implementation of the test
joincount.test(unemployment, listw2U(nb2listw(menir75.nb)))
```

---

Where dealing with several categories, the `joincount.multi` function of package `spdep` tests the significance, in accordance with the same principle, namely of spatial association of different variables.

■

## 3.3 Local measures of spatial autocorrelation

Global statistics are based on **the assumption of a spatial stationary process**: spatial autocorrelation would be the same throughout space. However, this assumption is all the less realistic as the number of observations is high.

### 3.3.1 Getis and Ord index

Getis and Ord (Getis et al. 1992) offer an indicator for identifying local spatial dependencies that do not appear in the global analysis.

### Getis and Ord indicator

$$G_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}} \quad (3.9)$$

$G_i > 0$  indicates a grouping of values higher than average.

$G_i < 0$  indicates a grouping of values lower than average.

The significance of the Getis and Ord indicator can be tested by making the assumption, in the absence of local spatial dependency, of a normal distribution.

$$z(G_i) = \frac{G_i - E(G_i)}{\sqrt{Var(G_i)}} \sim \mathcal{N}(0, 1) \quad (3.10)$$

### Application with R

The `localG` function of package `spdep` makes it possible to use this indicator.

#### 3.3.2 Local spatial autocorrelation indicators

Anselin (Anselin 1995) develops the concepts introduced by Getis and Ord by defining *local spatial autocorrelation indicators*. These must measure the intensity and significance of local autocorrelation between the value of a variable in a spatial unit and the value of the same variable in the surrounding spatial units. More specifically, these indicators make it possible to:

- detect significant groupings of identical values around a particular location (clusters);
- identify spatial non-stationarity zones, which do not follow the global process.

The Getis and Ord indicators serve only the first of these two objectives. To be considered as local spatial association measures – (LISA; *Local Indicators of Spatial Association*) – as defined by Anselin, these indicators must verify the following two properties:

- for each observation, they indicate the intensity of the grouping of similar – or opposite in trend – values around this observation;
- the sum of local indices on all observations is proportional to the corresponding global index.

One of the most used LISA is the local Moran's I.

### Local Moran's I

$$I_i = (y_i - \bar{y}) \sum_j w_{ij} (y_j - \bar{y}) \quad (3.11)$$

$$I_W = constante * \sum_i I_i \quad (3.12)$$

$I_i > 0$  indicates a grouping of similar values (higher or lower than average).

$I_i < 0$  indicates a combination of dissimilar values (e.g. high values surrounded by low values).

#### 3.3.3 Significance of the local Moran's I

Significant LISAs are combinations of similar or dissimilar values more marked than what might have been observed based on random spatial distribution. These groupings can match up with the four types of spatial groupings described in 3.1 and identifiable on Moran's diagram (high-high, low-low, high-low or low-low). The significance test of each local association indicator is based on

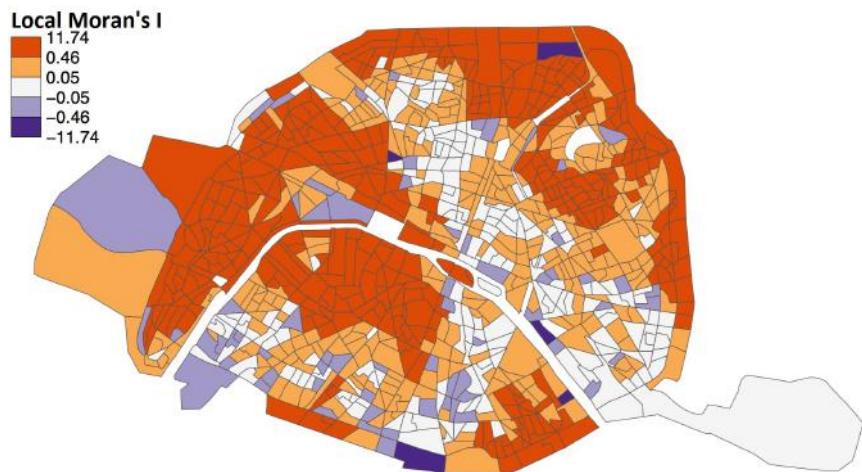


Figure 3.8 – Values of local Moran's I, on Parisian IRIS  
**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

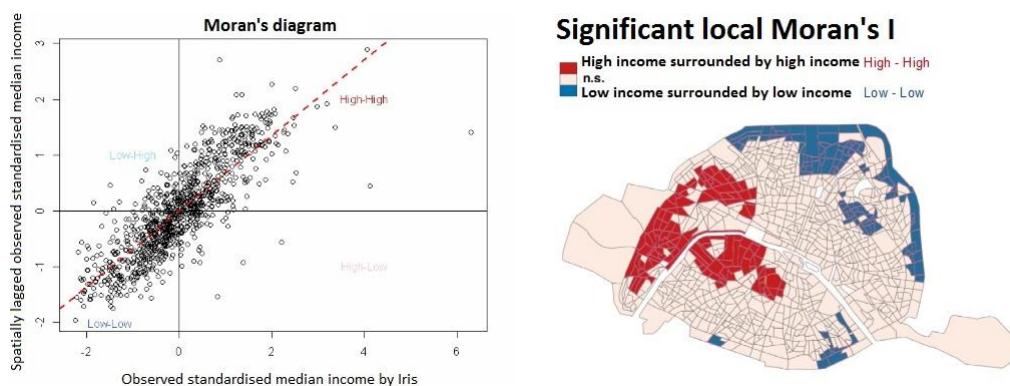


Figure 3.9 – Significant local Moran's I, on Parisian IRIS  
**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

a statistic assumed to asymptotically follow a normal distribution under the null hypothesis. If the assumption of normality holds,  $z(I_i) = \frac{I_i - E(I_i)}{\sqrt{Var(I_i)}} \sim \mathcal{N}(0, 1)$ .

To test the validity of the normality assumption of the LISAs under the null hypothesis, several random distributions are simulated in the space of the variable of interest and the local indicators associated with these simulations are calculated.

Taking up the example of Parisians' income once again, we can see that (Figure 3.10) the extreme quantiles of the distribution of the local Is are higher than those of a normal distribution. The *p-values* calculated under the normality assumption are therefore to be used with caution. This is because, as Anselin (Anselin 1995) shows, based on simulations (Figure 3.11), **in the presence of global spatial autocorrelation, the normality assumption of  $I_i$  no longer holds**.

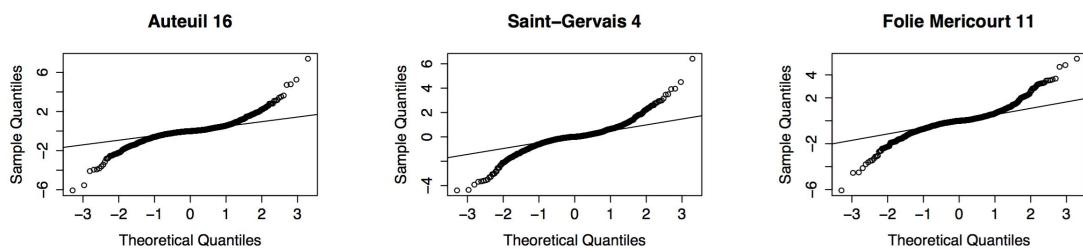


Figure 3.10 – Testing the normality assumption of local Moran's Is distribution on three Parisian IRIS

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

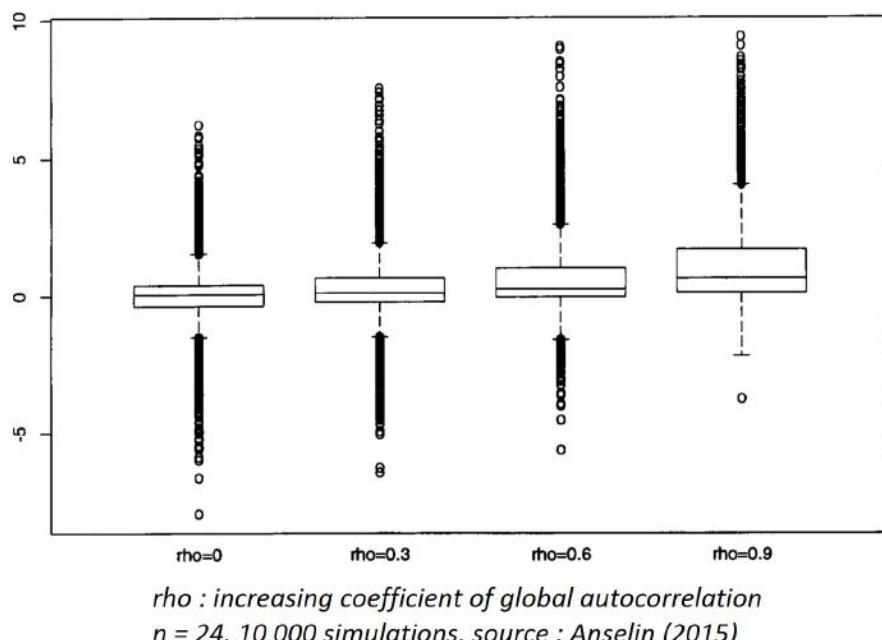


Figure 3.11 – Distribution of local Moran's Is where global spatial autocorrelation exists

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

Moreover, the LISAs significance test raises a problem encountered every time multiple comparisons are made. Indeed, when several statistical tests are carried out simultaneously using the same dataset, the global risk of errors in decision of the first kind – probability of wrongly

rejecting the null hypothesis – increases. At each test, the risk of randomly having a significant result is repeated. This increases the global risk to wrongly accept the significance of local index. Thus, in our case, we will conclude positively as to the **existence** of local spatial autocorrelation if **at least one** local spatial autocorrelation index is significant out of all indices in the study area. If there are 100 local spatial autocorrelation indices, there is a 100-fold increase in the risk of incorrectly detecting at least one significant one (exact formula in box 3.3.1). Risk inflation  $\alpha$  (type I error) is the risk of concluding incorrectly that local spatial autocorrelation has occurred is increased (Anselin 1995, Ord et al. 1995).

Different methods have been developed to prevent risk inflation  $\alpha$  when multiple statistical comparisons are needed. Some of them are described below. Let  $\alpha$  be the significance level selected for each local index.

**Box 3.3.1 — The Bonferroni method – the historical method.** The probability of not wrongly rejecting  $H_0$  is  $1 - \alpha$  by polygon, therefore  $(1 - \alpha)^n$  for the whole zone, with  $n$  the number of polygons.

The probability of rejecting  $H_0$  wrongly at least once is  $\alpha^* = 1 - (1 - \alpha)^n \approx n\alpha$ .

If the overall risk is to be maintained at approximately  $\alpha$ , it is thus possible to choose  $\alpha' \approx \frac{\alpha^*}{n}$  as a level for each individual test. For example, for  $\alpha = 0.05$ , a grouping is significant if its p-value is  $\frac{0.05}{n}$ .

The R software allows this method to be applied with the `method='bonferroni'` option of the `p.adjust` function.

It is considered that this method only yields good results when the number of tests carried out is small. In the case of the LISAs, it is a bit too restrictive and can lead to risk, given the number of comparisons, not to detect certain significant LISAs.

### Box 3.3.2 — The Holm Adjustment Method makes it possible to detect a spatial cluster.

The Holm adjustment method (Holm 1979) takes into account the fact that out of  $n$  polygons,  $k$  are truly significant spatial clusters, thus the probability of incorrectly rejecting  $H_0$  on the whole area is not  $(1 - \alpha)^n$  but  $(1 - \alpha)^{n-k}$ , where  $\alpha$  is the desired significance level.

The Holm method classifies the p-values from  $\alpha_1$  lowest to  $\alpha_n$  highest. If  $\alpha'_1 \sim n\alpha_1 < \alpha$ , i.e.  $\alpha_1 < \frac{\alpha}{n}$ , it is considered that this local index is indeed significant since it meets the most restrictive criterion. Attention is then turned to whether  $\alpha_2 < \frac{\alpha}{n-1}$ , and so on until testing whether  $\alpha_k < \frac{\alpha}{n-k+1}$ .

The R software makes it possible to apply this method with the `method='holm'` option of the `p.adjust` function.

The Holm adjustment method leads to more significant clusters than the Bonferroni method. It is therefore the most often preferred. However, this method also focuses on detecting **at least one cluster throughout the zone**.

**Box 3.3.3 — The False Discovery Rate Method – locating spatial clusters.** The False Discovery Rate (FDR) method was introduced by Benjamini et al. 1995. With this method, the risk of judging – incorrectly – a cluster as significant is higher, but conversely the risk of judging – incorrectly – a cluster as non-significant is lower. Caldas de Castro et al. 2006 prove the interest of this method to **locate** significant spatial clusters.

The FDR method classifies the p-values from  $\alpha_1$  lowest to  $\alpha_n$  highest.

Let  $k$  be the largest whole number such that  $\alpha_k \leq \frac{k}{n}\alpha$ . Benjamini and Hochberg explain that the null hypothesis of absence of local spatial autocorrelation for all clusters whose p-values are less than or equal to  $\alpha_k$  can be rejected. The R software makes it possible to apply this method with

the method='fdr' option of the p.adjust function.

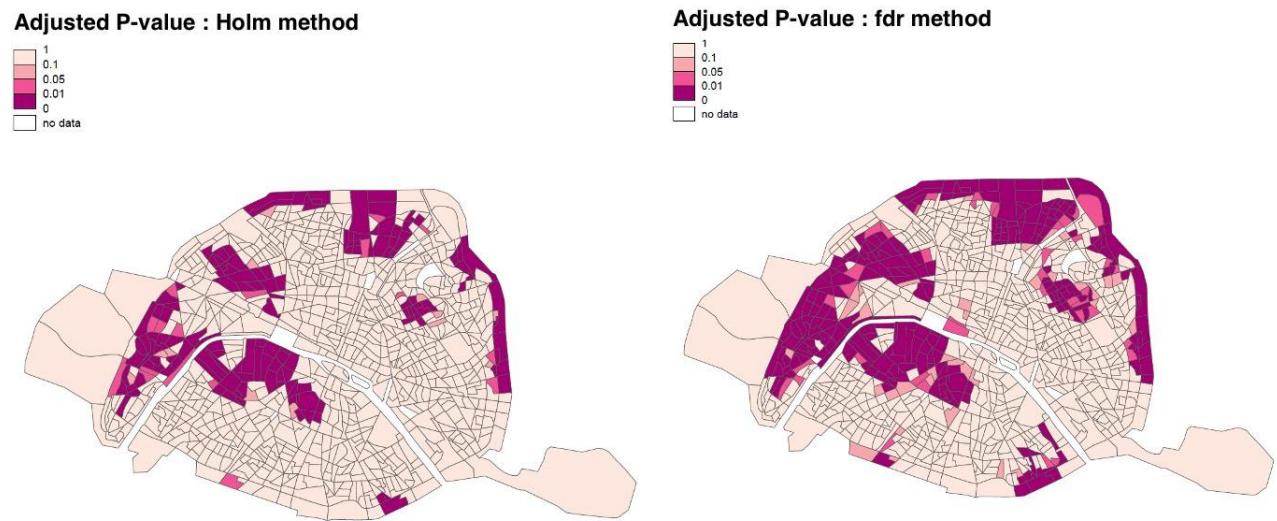


Figure 3.12 – Testing the significance of the local Moran's I, on Parisian IRIS

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

In the example of Parisians income (Figure 3.12), it is clear that the adjustment of the p-values using the Holm method leads to less significant p-values than the adjustment by the FDR method. The Holm method reduces the risk of making incorrect positive conclusions as to the **existence** of local spatial autocorrelation. On the other hand, this method increases the risk of overlooking a local cluster. The choice of the adjustment method will therefore depend on the objectives of the study and the risks that are favoured.

#### Application with R

```
lisa_revenus<- localmoran(iris75.data$med_revenu, iris75.lw, zero.policy=TRUE)

# Calculation of adjusted p-values
iris75.data.LISA$pvalue_ajuste<-
p.adjust(iris75.data.LISA$pvalue_LISA, method='bonferroni')
```

### 3.3.4 Interpretation of local indices

#### In the absence of global spatial autocorrelation

The LISAs make it possible to **identify areas where similar values are significantly grouped**. These are areas where the local spatial structure is such that the relations between neighbours are particularly strong.

#### In the presence of global spatial autocorrelation

LISAs **indicate areas that have a particular impact on the global process** (local autocorrelation more pronounced than global autocorrelation), **or, on the contrary, which stand out from it** (lower autocorrelation). Thus, using the median income of Parisians as an example, it can be seen that the distribution of local Moran's I is not centred on the global Moran's I (Figure 3.13). Some zones have a significantly different spatial association structure than the global process.

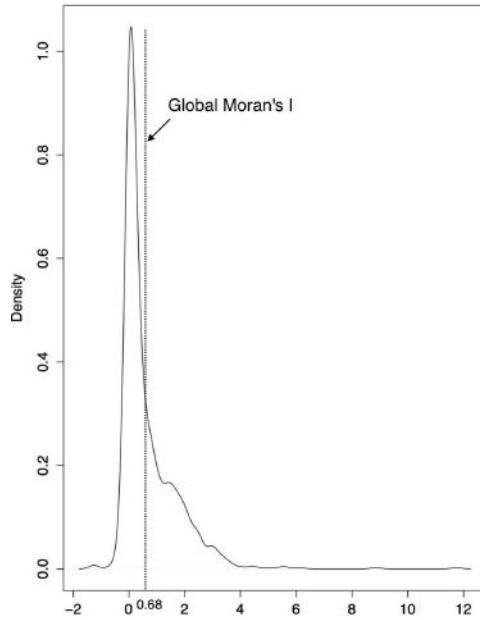


Figure 3.13 – Distribution of local Moran's I of median incomes by Parisian IRIS

**Source:** INSEE, Localised Tax Revenues System (RFL) 2010

Even adjusted, the p-values are at risk of being too low, as the distribution of  $I_i$  moves away from the norm. The more the global autocorrelation increases, the higher the number of extreme values. High LISAs can therefore hardly be interpreted as significant groupings of similar value. In this case, LISAs are interpreted as indicators of a certain type of **local instability**.

### 3.4 Spatio-temporal indices

It is not unusual for a geolocalised database to have observations raised at different points in time, as is the case with databases that list real estate transactions. It may be interesting to understand how a localised phenomenon has spread and evolved in space and time and how it can be linked to the conditions of the environment surrounding it. In this case, it is important to be able to assess how the underlying spatial structures change over different periods of time. On spatio-temporal data, prior graphical exploration of cross-section data (standard Moran's I) can be used to study the existence and change in grouping or dispersion trends that are statistically significantly different from random models. Many recent developments show a growing interest in analysing spatio-temporal data in many areas of research such as physics, meteorology, economics and environmental studies. By extending the Moran index to include time attributes, it becomes possible to calculate global and localized indices that concurrently take into account spatial and temporal auto-correlations. This can also be done on the basis of spatial-temporal risk weighting matrices. The work of Martin et al. 1975, Wang et al. n.d., López-Hernández et al. 2007 suggests extensions of Moran's I, traditionally used to measure spatial dependence, to calculate a spatio-temporal Moran's I. Chen et al. 2013 develop an enhanced analytical approach based on the traditional Moran's I, based on stationary data over time. As Lee et al. 2017 note, geolocated time series are usually non-stationary. When this assumption is not respected, Moran's spatiotemporal index suggested by Chen et al. 2013 can be fallacious. Lee et al. 2017 suggest to bypass this difficulty by applying a correction of fluctuations around the trend – detrended fluctuation analysis, DFA – and suggest a new method for calculating this index.

## Conclusion

Spatial autocorrelation indices are exploratory statistical tools that make it possible to bring out the existence of a significant spatial phenomenon. Sections 2 and 3 present different methods of taking this spatial phenomenon into account, at global or local level, for quantitative or qualitative variables. It is important to know whether the autocorrelation is insignificant, but also to measure the extent to which the autocorrelation is significant in order to determine the scale of spatial dependence. The study of spatial autocorrelation is an essential step before considering any specification of spatial interactions in an appropriate model.

### References - Chapter 3

- Anselin, Luc (1995). « Local indicators of spatial association—LISA ». *Geographical analysis* 27.2, pp. 93–115.
- Benjamini, Yoav and Yosef Hochberg (1995). « Controlling the false discovery rate: a practical and powerful approach to multiple testing ». *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.
- Bivand, Roger S, Edzer Pebesma, and Virgilio Gomez-Rubio (2013a). *Applied spatial data analysis with R*. Vol. 10. Springer Science & Business Media.
- Caldas de Castro, Marcia and Burton H Singer (2006). « Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association ». *Geographical Analysis* 38.2, pp. 180–208.
- Chen, Shao-Kuan et al. (2013). « Analysis on Urban Traffic Status Based on Improved Spatiotemporal Moran's I ». *Acta Physica Sinica* 62.14.
- Getis, Arthur and J Keith Ord (1992). « The analysis of spatial association by use of distance statistics ». *Geographical analysis* 24.3, pp. 189–206.
- Holm, Sture (1979). « A simple sequentially rejective multiple test procedure ». *Scandinavian journal of statistics*, pp. 65–70.
- Lee, Jay and Shengwen Li (2017). « Extending moran's index for measuring spatiotemporal clustering of geographic events ». *Geographical Analysis* 49.1, pp. 36–57.
- López-Hernández, Fernando A and Coro Chasco-Yrigoyen (2007). « Time-trend in spatial dependence: Specification strategy in the first-order spatial autoregressive model ». *Estudios de Economía Aplicada* 25.2.
- Martin, Russell L and JE Oeppen (1975). « The identification of regional forecasting models using space: time correlation functions ». *Transactions of the Institute of British Geographers*, pp. 95–118.
- Openshaw, Stan (1984). *The modifiable areal unit problem*. Vol. CATMOG 38. GeoBooks, Norwich, England.
- Openshaw, Stan and Peter Taylor (1979b). « A million or so correlation coefficients ». *Statistical methods in the spatial sciences*, pp. 127–144.
- Ord, J Keith and Arthur Getis (1995). « Local spatial autocorrelation statistics: distributional issues and an application ». *Geographical analysis* 27.4, pp. 286–306.
- Tiefelsdorf, Michael (1998). « Modelling spatial processes: The identification and analysis of spatial relationships in regression residuals by means of Moran's I (Germany) ». PhD thesis. Université Wilfrid Laurier.
- Upton, Graham, Bernard Fingleton, et al. (1985). *Spatial data analysis by example. Volume 1: Point pattern and quantitative data*. John W & Sons Ltd.
- Wang, Y. F. and H. L. He. « Spatial Data Analysis Method ». *Science Press, Beijing, China*.
- Zhukov, Yuri M (2010). « Applied spatial statistics in R, Section 2 ». *Geostatistics.[Online]* Available: <http://www.people.fas.harvard.edu/~zhukov/Spatial5.pdf>.

# 4. Spatial distribution of points

JEAN-MICHEL FLOCH

INSEE

ERIC MARCON

*AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana.*

FLORENCE PUECH

*RITM, Univ. Paris-Sud, Université Paris-Saclay & CREST, 92330 Sceaux, France.*

---

<b>4.1</b>	<b>Framework of analysis: basic concepts</b>	<b>74</b>
4.1.1	Configurations and processes .....	74
4.1.2	Marked processes .....	75
4.1.3	Observation window .....	75
<b>4.2</b>	<b>Point processes: a brief presentation</b>	<b>76</b>
4.2.1	The homogeneous Poisson process .....	76
4.2.2	Intensity, first-order property .....	78
4.2.3	The Inhomogeneous Poisson Process .....	79
4.2.4	Second-order properties .....	79
<b>4.3</b>	<b>From point processes to observed point distributions</b>	<b>81</b>
4.3.1	Distribution by random, aggregation, regularity .....	81
4.3.2	Warnings .....	82
<b>4.4</b>	<b>What statistical tools should be used to study spatial distributions?</b>	<b>83</b>
4.4.1	Ripley's $K$ function and its variants .....	83
4.4.2	How can we test the significance of the results? .....	88
4.4.3	Review and focus on important properties for new measurements ..	90
<b>4.5</b>	<b>Recently proposed distance-based measures</b>	<b>95</b>
4.5.1	The $K_d$ indicator of Duranton and Overman .....	95
4.5.2	$M$ function of Marcon and Puech .....	96
4.5.3	Other developments .....	98
<b>4.6</b>	<b>Multi-type processes</b>	<b>98</b>
4.6.1	Intensity functions .....	98
4.6.2	Intertype functions .....	102
<b>4.7</b>	<b>Process modelling</b>	<b>107</b>
4.7.1	General modelling framework .....	107
4.7.2	Application examples .....	107

---

## Abstract

Statisticians carry out close examination of spatialized data, such as the distribution of household income, the location of industrial or commercial establishments, the distribution of schools in cities, etc. Answers can be found through analyses of one or more predefined geographical scales such as neighbourhoods, districts or statistical blocks. However, it is tempting to preserve the individual data and to work with the exact position of the entities that are being studied. If that is the case, statisticians have to conduct analyses based on geolocation data without carrying out any geographical aggregation. Observations are taken as points in space and the objective is to characterise these point distributions.

Understanding and mastering statistical methods that process this individual and spatialized information enables us to work on data that are now increasingly accessible and sought after because they provide very precise analyses of distributions studied (Ellison et al. 2010; Barlet et al. 2013). In this framework of analysis, statisticians who have sets of points to analyse are faced with several important methodological questions: how can such data with thousands or even millions of observations be represented and characterised spatially? What statistical tools exist that can be used to study these observations relating to households, employees, firms, stores, equipment or travel, for example? How can the qualitative or quantitative characteristics of the observations being studied be taken into account? How can any attractions or repulsions between points or between different types of points be highlighted? How can we assess the significance of the results obtained, etc?

The purpose of this chapter is to help statisticians to provide statistically robust results from the study of spatialized data that is not based on predefined zoning. To do this, we will review the literature on the subject of statistical methods used to characterise point distributions and we will explain the associated issues. We will use simple examples to explain the advantages and disadvantages of the most frequently adopted approaches. The code provided in R will be used to reproduce the examples covered.

**Acknowledgements:** The authors would like to thank Gabriel LANG and Salima BOUAYAD AGHA for their careful review of the first version of this chapter and for all their constructive comments. Thanks also to Marie-Pierre de BELLEFON and Vincent LOONIS who provided the initiative for this project: this chapter has undeniably benefited from all their editorial efforts and those of Vianney COSTEMALLE.

## Introduction

The study of spatial distributions of points may seem more removed from the concerns of public statisticians than some other methods. So why give them a place in this manual? The answer is simple: geolocation of data provides numerous localised observations on firms, facilities and housing. This swiftly leads us to consider the possibility of gathering together these observations, the spatial configuration of their random, or non-random setting, and their dependence on other processes (the proximity of industrial establishments with strong *input-output* links may be desirable and therefore lead to spatial interactions between establishments from different sectors). The aim of this chapter is to present an introduction to a body of methods that are sometimes complex in their mathematical foundations, but which often serve to illustrate quite simple questions. The development of these methods was based in the issues facing ecologists, foresters and epidemiologists. P.J. Diggle, the author of the first reference work (Diggle 1983), is known for his extensive epidemiology work (Diggle et al. 1991). As a result, educational examples illustrating point processes often come from forestry or epidemiological data. In this chapter we will use examples of this type provided in certain R packages such as *spatstat* (Baddeley et al. 2005) or *dbmss* (Marcon et al. 2015b). We will also use data on the location of facilities in France.

Unlike zoning or geostatistical methods, when studying spatial distribution, a variable is not measured locally, but the very location of the points is at the heart of the subject in question. We will build models and make inferences based on these points.

The maps in Figure 4.1, produced from data in the permanent database of facilities (BPE), show four examples of the location of activities in the city of Rennes (France).<sup>1</sup>

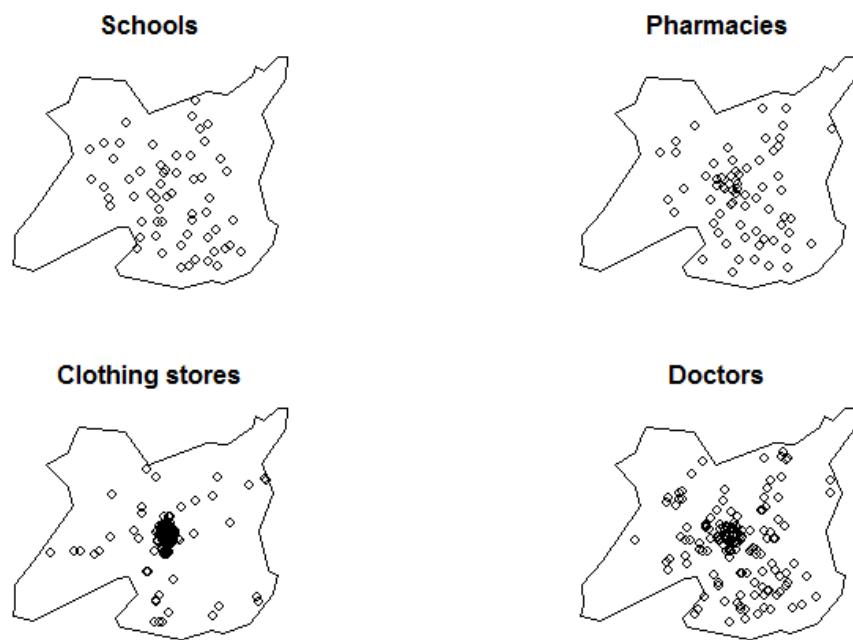


Figure 4.1 – Four examples of the location of activities in the municipality of Rennes in 2015  
**Source:** INSEE-BPE, authors' calculations

1. If equipment is positioned imprecisely, it is assigned by default to the centroid of the associated IRIS (INSEE zoning in "Ilots Regroupés pour l'Information Statistique" that can be translated as "aggregated units for statistical information", see <https://www.insee.fr/en/metadonnees/definition/c1523>).

---

```

library("spatstat")
library("sp")
# BPE file on the INSEE.fr site: https://www.insee.fr
# Data for these examples:
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))
bpe_sch <- bpe[bpe$TYPEQU=="C104", ]
bpe_pha <- bpe[bpe$TYPEQU=="D301", ]
bpe_clo <- bpe[bpe$TYPEQU=="B302", ]
bpe_doc <- bpe[bpe$TYPEQU=="D201", ]
par(mfrow=c(2,2), mar=c(2, 2, 2, 2))
plot(carte, main="Schools") ; points(bpe_sch[, 2:3])
plot(carte, main="Pharmacies") ; points(bpe_pha[, 2:3])
plot(carte, main="Clothing stores") ; points(bpe_clo[, 2:3])
plot(carte, main="Doctors") ; points(bpe_doc[, 2:3])
par(mfrow=c(1,1))

```

---

These four simple figures provide an initial overview of the major differences in the locations of these facilities. There is a large number of clothing stores, but they are extremely concentrated in the center of Rennes. On the other hand, primary schools seem to be distributed more evenly. Pharmacies are also evenly distributed, but with a greater presence in the city center. The location of doctors is more aggregated than that of pharmacies, but less so than that of clothing stores. These initial conclusions on the distribution of activities could be supplemented by more advanced spatial analyses, for example by applying data for population distribution or accessibility (closer to or further away from the main communication routes). The methods presented in this manual make it possible to go beyond the conclusions of these first maps, which are certainly informative but insufficient to characterise and explain the location of the entities in question.

In this chapter, we have chosen not to deal with methods that discretise space, *i.e.* approaches based on study zoning (such as employment zones in France based on commuting patterns) or administrative zoning (such as the breakdown of the Nomenclature of Territorial Units for Statistics - NUTS - from Eurostat). Specific works (Combes et al. 2008) provide a very good introduction to this subject for any interested readers. This chapter will be limited to methods that take into account the exact geographical position of the entities studied. Our choice is motivated by at least two factors. The first is linked to access to such data on a large scale and the development of appropriate technical methods to analyse them in a meaningful way. Different packages are, for example, accessible in the R software. The second is that by favouring methods that preserve the nature of the individual data analysed (position in space, characteristics), the Modifiable Areal Unit Problem - MAUP, well known to geographers (Openshaw et al. 1979a), will be avoided. MAUP refers to the fact that the discretisation of initially non-aggregated data potentially creates several statistical biases linked to the position of borders, aggregation level etc. (Briant et al. 2010).

## 4.1 Framework of analysis: basic concepts

This section aims to define the fundamental concepts we will use in this chapter to explain statistical methods of spatial analysis of point data.

### 4.1.1 Configurations and processes

To study these empirical **spatial distribution** of points (or set of points), we use the random point process theory. A point process can be used to randomly generate an infinity of outcomes, which share a number of properties.

Usually, we note the point process as  $X$  and a realization from this process as  $S$ . Spatial distributions are modelled using inferential methods that apply to objects that are observed only once. For example, for many data, statisticians only have one set of points observed at a given date. Therefore, there is only one distribution of doctors in the city of Rennes (see figure 4.1), bus stops in London, housing in Friesland in the Netherlands or cinemas in Belgium on a given date. However, the unique observed realization must not alter our analysis: we will, therefore, ensure that the available data is able to provide a good approximation of the point process that generated it. We will come back to this in this chapter.

**Definition 4.1.1 — Spatial distribution.** A distribution of  $n$  points, written  $C = \{x_1, \dots, x_n\}$  is a set of points from  $\mathbb{R}^2$  in this chapter: the objects are located on a map. The theory does not limit the dimension of space but applications in three-dimensional spaces are rare, and almost non-existent in  $\mathbb{R}^d$ ,  $d > 3$ . The number of points in the distribution is noted as  $n(C)$ . The points are not considered to be duplicated, as this would prevent many methods from being used. **The combined points in the region  $B$  is written  $C \cap B$ , and  $n(C \cap B)$ .**

The process  $X$  is defined if the number of points  $n(X \cap B)$  is known for any region  $B$ . The number of points is also written  $N(B)$  if no confusion is possible. In general, we are limited to locally finite processes, for which  $n(X \cap B) < +\infty, \forall B$ .

#### 4.1.2 Marked processes

One or more characteristics can be associated with each point. These characteristics are known as marked points. In this case, we talk about **marked point processes**. This approach has been widely used in forest studies (see for example Marcon et al. 2012).

The marks used can be qualitative (different tree species) or quantitative (trunk diameter, tree size). If we take the example of clothing shops, qualitative markers could be the type of store (ready-to-wear or made-to-measure) and quantitative marks could be store surface area or number of employees. Marks can be more sophisticated. For example, Florent Bonneau characterised the spatial distribution of incidents in the Toulouse region in 2004 using the associated workload for each fire service intervention (Bonneu 2007). This quantitative mark is obtained by multiplying the duration of the intervention and the number of firefighters mobilised.

To begin, we will limit ourselves to unmarked processes.

#### 4.1.3 Observation window

The area to study the location of points is often called the **window** and it is often arbitrary. The authors take an area for study that may be square (Møller et al. 2014), rectangular (Cole et al. 1999), circular (Szwagrzyk et al. 1993), an administrative area (Arbia et al. 2012) or study zone (Lagache et al. 2013).

The indicators used to detect the underlying spatial structures are based on an analysis of the **neighbourhood of points**: for example, for all the points studied, the average number of similar points within a radius of 2 km, 4 km, etc. It may then be necessary to take into account points located on the edge of the area of interest. The risk is to underestimate the neighbourhood of points located on the edge of the area, as some of their neighbours are located outside the area. For example, we can see this in figure 4.2. Let us assume that the area being studied is a square plot within a forest and that the points represent trees. The neighbourhood of points  $i$  is described as the circle with a radius  $r$ , centred on point  $i$ . If you want to estimate the number of neighbours for point  $i$ , counting only the points in the circle that are included within the parcel would underestimate the actual number of its neighbours. The reason is simple: a part of the circle is located outside the field of study.

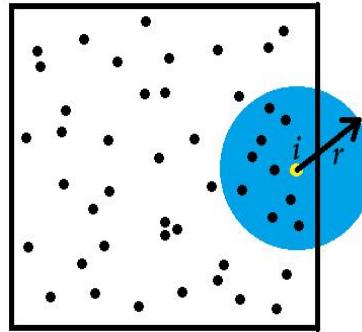


Figure 4.2 – Edge effect example

**Source:** the authors

The study by Marcon et al. 2003 illustrates, for example, the importance of not taking this bias into account when estimating the concentration of industrial activities in France. Generally, regardless of the area of application, this potential bias is deemed severe enough for the use of a corrective technique to account for “edge effects”. There is a great deal of literature on these edge effects and their correction (overall or individual correction, creation of a buffer zone around the area, use of toroidal correction<sup>2</sup>...) Interested readers may refer to traditional spatial statistics manuals for further information (Illian et al. 2008 ; Baddeley et al. 2015b). From a practical point of view, calculation software (and in particular R) can be used to treat these effects using different correction methods. An example will be provided in chapter 8: "Spatial smoothing".

## 4.2 Point processes: a brief presentation

### 4.2.1 The homogeneous Poisson process

To begin, let's look at the point process that is used to generate completely random spatial point distributions (Complete Spatial Randomness - CSR). To achieve this, we can start with a particularly simple process,  $U$ , which generates a single point that can be randomly located in an area of interest  $W$ . If  $u_1$  and  $u_2$  are the coordinates of the point, it is possible to calculate the probability that the point generated by  $U$  is located in a small space  $B$ , which is selected arbitrarily:

$$P(U \in B) = \int_B f(u_1, u_2) du_1 du_2. \quad (4.1)$$

The distribution is uniform over  $W$  iff  $f(u_1, u_2) = \frac{1}{|W|}$  where  $|W|$  designates the area of  $W$ .

Therefore, we have:  $P(U \in B) = \int_B f(u_1, u_2) du_1 du_2 = \frac{1}{|W|} \int_B du_1 du_2 = \frac{|B|}{|W|}$ . This process allows another process to be defined - the binomial process.  $n$  points are distributed evenly across the region  $W$ , independently. Traditionally, we would write that:

$$P(n(X \cap B) = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (4.2)$$

with  $p = \frac{|B|}{|W|}$ . The `runitpoint` function in the R package `spatstat` generates spatial distributions of points from a uniform binomial process. For example, in figure 4.3, 1,000 points are expected in a 10 x 10 observation window.

2. The toroidal correction can be applied to a rectangular window. The window is folded over onto itself to form a torus: continuity is established between the right and left limits (upper and lower, respectively) of the window, which, therefore, no longer has any edge

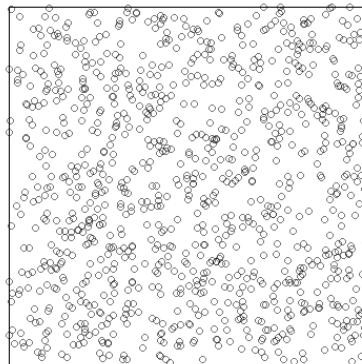


Figure 4.3 – 1 000 points sample using a uniform binomial process

**Source:** package *spatstat*, authors' calculations

---

```
library("spatstat")
plot(runifpoint(1000, win=owin(c(0, 10),c(0, 10))), main="")
```

---

Why is such a process, in which each point is placed uniformly at random, not appropriate to define a CSR process? Initially, we require two properties from such a process:

- **homogeneity** which corresponds to the absence of “preference” for a particular location (this is indeed the case for the binomial process).
- **Independence**, to reflect the fact that realizations in one area of the space have no influence on realizations in another region. This is not the case for the binomial process.

If there are  $k$  points in the  $B$  area of  $W$ , there are  $n - k$  in the rest of the area.

Homogeneity induces that the number of points expected in the  $B$  region is either proportional to its surface, or  $E [n(X \cap B)] = \lambda |B|$ .  $\lambda$  is a constant that corresponds to the average number of points per unit of surface area. The Poisson law, which will be used to characterise a CSR process, can be introduced heuristically based on the property of independence. This implies that all counts in grids are independent, regardless of the size of the square. When cells, numbered  $m$ , become extremely small, most of them contain no points and some contain only one. The probability of a region containing more than one point becomes negligible. Based on the hypothesis of independence,  $n(X \cap B)$  is the number of successes from a large number of independent drawings, with each drawing having a very low probability of success. This number of successes follows a binomial law of parameters  $m$  and  $\lambda |B| / m$ , which tends towards the Poisson law for the  $\lambda |B|$  parameter when  $m$  becomes large:

$$P(n(X \cap B) = k) = e^{-\lambda |B|} \frac{\lambda^k |B|^k}{k!}. \quad (4.3)$$

Therefore, we come to this conclusion on the basis of the hypotheses of homogeneity and independence.

**Definition 4.2.1 — CSR process.** The CSR process or homogeneous Poisson process is often defined as follows:

- $P(n(X \cap B) = k) = e^{-\lambda|B|} \frac{\lambda^k |B|^k}{k!}$ .  
This defines the Poissonian nature of the distribution (**PP1**);
- $E[n(X \cap B)] = \lambda |B|$ .  
This defines the homogeneity (**PP2**);
- $n(X \cap B_1), \dots, n(X \cap B_m)$  are  $m$  independent random variables (**PP3**);
- once the number of points is set, the distribution is uniform (**PP4**).

Properties **PP2** and **PP3** are sufficient to define the CSR process (Diggle 1983), and it can be demonstrated that others are consequential. Other properties result from this. Firstly, the superposition of independent Poisson processes with parameters  $\lambda_1$  and  $\lambda_2$  gives a Poisson process with a parameter of  $\lambda_1 + \lambda_2$ . If points are eliminated randomly with a constant probability  $p$  in a Poisson process (*thinned process*), the resulting process is always a Poisson process with parameter  $p\lambda$ , where  $p$  is the thinning parameter.

The homogeneous Poisson process plays a decisive role in modelling spatial distributions of points<sup>3</sup>. Many spatial processes have been defined, and we will give a few examples in this chapter. These can be implemented using package *spatstat*. For example, the *rpoispp* function will be used to simulate homogeneous Poisson processes. Figure 4.4 is a realization of a homogeneous Poisson process in a  $1 \times 1$  observation window: 50 points are expected and the points are distributed completely randomly over the window.

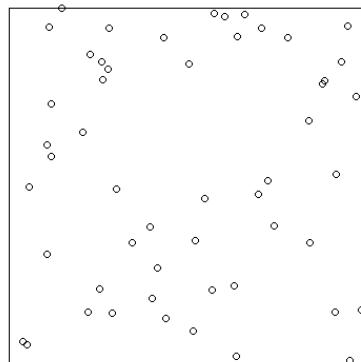


Figure 4.4 – 50 points sample by a homogeneous Poisson process

**Source:** package *spatstat*, authors' calculations

---

```
library("spatstat")
plot(rpoispp(50), main="")
```

---

#### 4.2.2 Intensity, first-order property

Process laws are very complex (Møller et al. 2004), which in practice leads to the preferred use of indicators that are qualified as first-order or second-order, in the same way as first-order and second-order moments (expectation and variance) are used to identify a random variable of unknown law.

3. A little like the Normal law in classical inferential statistics (although its properties make it closer to the uniform law).

**Definition 4.2.2 — Intensity of a process.** Intensity featured in the presentation of the Poisson process, where it was constant ( $\lambda$ ). There are other processes in which this hypothesis is rejected, and in which the intensity function  $\lambda(x)$  is variable. It is defined as  $E[n(X \cap B)] = \mu(B) = \int_B \lambda(x) dx$ .

By applying the definition of expectation to a small region centred on  $x$  and surface  $dx$ , we can define **the intensity** at this point  $x$  as **the number of points expected in this small area when it tends towards 0**, or:

$$\lambda(x) = \lim_{|dx| \rightarrow 0} \frac{E[N(dx)]}{|dx|}. \quad (4.4)$$

If it is not constant, it may be **estimated using non-parametric methods** that are used for density estimation. In its simplest version, without correcting edge effects, the intensity estimator is written:  $\hat{\lambda}(u) = \sum_{i=1}^n K(u - x_i)$ ,  $K$  designating the kernel, which can be Gaussian, or with finished support (Epanechnikov kernel, Tukey's biweight kernel). They must check that  $\int_{\mathbb{R}^2} K(u) du = 1$ . As in all non-parametric methods, **the choice of kernel has a limited impact**. In contrast, **the choice of bandwidth is extremely important** (see, for example Illian et al. 2008). A presentation of these estimation methods can be found in chapter 8 of this manual: "Spatial smoothing". The function used in the R software is `density` in package `spatstat`, which provides contours, 3D representations and colour degradations. Several examples will be given in section 4.6.1 of this chapter.

#### 4.2.3 The Inhomogeneous Poisson Process

Inhomogeneous Poisson processes are of variable intensity and their points are distributed independently of each other (the **PP3** condition is maintained). The **PP1** condition regarding the Poissonian nature of the distribution, conditional to  $n$ , is maintained, as the parameter for the law is no longer  $\lambda |B|$ , but  $\mu(B)$  as defined above. The **PP4** condition is modified. Subject to a number of fixed points  $n$ , the points are independent and identically distributed, with a probability density of  $f(x) = \frac{\lambda(x)}{\int_B f(u) du}$ .

Figure 4.5 shows two examples of inhomogeneous Poisson processes, characterised by their intensity function (with coordinates  $x$  and  $y$ ).

---

```
library("spatstat")
par(mfrow=c(1, 2))
plot(rpoispp(function(x, y) {500*(x+y)}), main=expression(lambda==500*(x+y)))
plot(rpoispp(function(x,y) {1000*exp(-(x^2+y^2)/.3)}), main=expression(
  lambda==1000*exp(-(x^2+y^2)/.3)))
par(mfrow=c(1,1))
```

---

#### 4.2.4 Second-order properties

To introduce the second-order properties of a point process, we will look at the **variance and covariance of point counts**, defined below:

$$var(n(X \cap B)) = E[n(X \cap B)^2] - E[n(X \cap B)]^2 \quad (4.5)$$

$$cov[n(X \cap B_1), n(X \cap B_2)] = E[n(X \cap B_1)n(X \cap B_2)] - E[n(X \cap B_1)]E[n(X \cap B_2)] \quad (4.6)$$

$$\lambda = 500(x+y)$$

$$\lambda = 1000 \exp(-(x^2+y^2)/0.3)$$

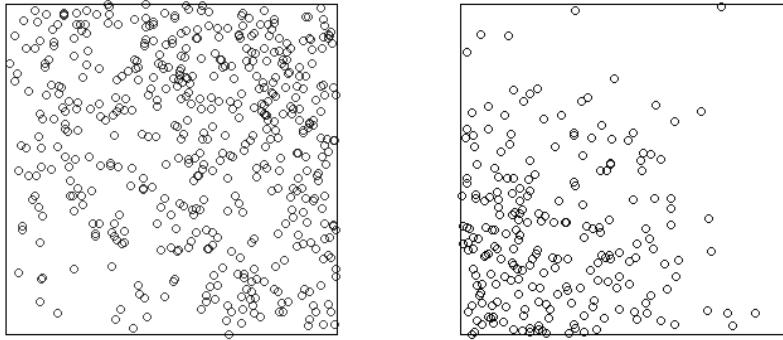


Figure 4.5 – Examples of inhomogeneous processes

**Source:** package *spatstat*, authors' calculations

**Definition 4.2.3 — Second-order moment of a process.** Rather than using these indicators, the second-order moment is defined as follows:

$$v_{|2|}(A \times B) = E[n(X \cap A)n(X \cap B)] - E[n(X \cap A \cap B)], \quad (4.7)$$

which, for the Poisson process, gives:  $\lambda^2 |A| |B|$ . When this measure includes a density, it is called order 2 intensity and noted  $\lambda_2$ . It is defined as  $v_{|2|}(C) = \int_C \lambda_2(u, v) du dv$ .

This second-order intensity can be interpreted as:

$$\lambda_2(x, y) = \lim_{|dx| \rightarrow 0, |dy| \rightarrow 0} \frac{E[N(dx)N(dy)]}{|dx| |dy|}. \quad (4.8)$$

First- and second-order intensities are used to define a function, called the *point pair correlation function*, as follows:

$$g_2(u, v) = \frac{\lambda_2(u, v)}{\lambda(u)\lambda(v)}. \quad (4.9)$$

In the case of a homogeneous Poisson process,  $\lambda_2(u, v) = \lambda^2$ ,  $g_2(u, v) = 1$ .

When a process is **stationary (at the second order)**<sup>4</sup>, the intensity of the second order is not affected by translation and depends only on the difference between the points:  $\lambda_2(x, y) = \lambda_2(x - y)$ .

When it is also **isotropic**, the process is not affected by rotation and the second-order intensity depends only on the distance between  $x$  and  $y$ . Note that second-order stationarity and isotropy are essential for many spatial statistical tools.

4. The term stationary, without any further details, is often used for constant order 1 and 2 intensity processes; first-order stationarity is synonymous with homogeneity.

## 4.3 From point processes to observed point distributions

### 4.3.1 Distribution by random, aggregation, regularity

When you look at a distribution of points, two main questions arise: are the observed points distributed randomly or is there an interaction? If there is interdependence, is it aggregate or repellent? Depending on the answers to these questions, **three spatial distributions** are generally found: a so-called completely random distribution, an aggregate and a regular distribution. An example of these three theoretical distributions is shown in figure 4.6. These spatial distributions are obtained from known point processes, simulated using package *spatstat*.

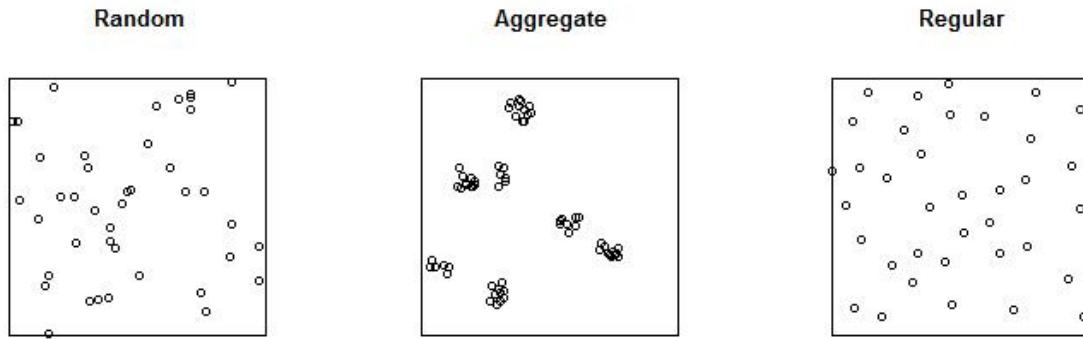


Figure 4.6 – The three standard spatial distributions of points

**Source:** package *spatstat*, authors' calculations

---

```
library("spatstat")
par(mfrow=c(1, 3))
plot(rpoispp(50), main="Random")
plot(rMatClust(5, 0.05, 10), main="Aggregate")
plot(rMaternII(200,0.1), main="Regular")
par(mfrow=c(1,1))
```

---

The **completely random** configuration is central to the theory. All spatial distributions, as point process realizations, are random but this corresponds to a “completely random” distribution of points on a surface: points are located everywhere with the same probability and independently of each other. This distribution corresponds to a realization of a homogeneous Poisson process. In this case, there is no interaction between the points but only the use of indicators makes it possible to judge whether the observed distribution differs *significantly* from a completely random distribution. Indeed, it is extremely difficult to identify such a configuration with the naked eye. In this example, we selected the *rpoispp* function in package *spatstat* to simulate the homogeneous Poisson processes.

The second distribution of points is said to be **regular**: consider the spatial distribution of trees in an orchard or along streets in town, the distribution of deckchairs on a beach, etc. In such a configuration, the points are *more regularly spaced* than they would be in a completely random distribution. Points repel each other and create a dispersed points distribution. A dispersion phenomenon can be seen for certain commercial activities, such as gas stations in Lyon (France, see Marcon et al. 2015a). Location constraints can also create dispersions, the geographic distribution of the capitol buildings in the USA is a good example of this (Holmes et al. 2004). In the right-hand chart of figure 4.6, we used a realization of a Matern process to represent a dispersed point distribution. Specifically, two simple examples of repellent processes are provided by the Matern I and II processes (see Baddeley et al. 2015b). In process I, all point pairs located at distances below

a threshold  $r$  are deleted. In process II, each point is marked by an arrival time, a random variable in  $[0, 1]$ . Points located at a lesser distance than  $r$  from a previously determined point are deleted. Using package *spatstat*, the *rMaternI* and *rMaternII* functions can be used to simulate these two Matern processes. In the example given in figure 4.6, we used a sample of a Matern type II process obtained using this package. It should be noted that other dispersed distributions can be observed: intuitively, for example, a dispersion phenomenon can be seen in a distribution of points located at the intersections of a honeycomb pattern: in this case the distance between the points is maximum (and it is greater than it would have been if the distribution was random).

Finally, the last possible configuration is known as **aggregated**. In this case, an interaction between the points can be seen. They attract each other, creating aggregates: a geographic concentration can then be detected. Looking at figure 4.1 in the introduction, it seems that the clothing stores in Rennes are mainly located in the city centre. This observation could be shared with other types of shops, such as clothing in specialised stores in the city Lyon (Marcon et al. 2015a). An aggregated configuration corresponds, for example, to the central theoretical case in figure 4.6 which is obtained by drawing a Matern cluster process. The idea of this process to simulate aggregates is quite intuitive. Around each "parent" point, in a circle with radius  $r$ , "offspring" points are distributed uniformly. In package *spatstat*, the *rMatClust* function can be used to simulate Matern cluster process realizations. We used this function to obtain the aggregated distribution in figure 4.6. In particular, we specified the intensity of the Poisson process for the parent points (equal to 5) and the average number of offspring points (10) drawn around the parent points in a circle of radius  $r$  (equal to 0.05).

### 4.3.2 Warnings

These spatial structures (aggregated, random or dispersed) are open to a very intuitive interpretation based on the hypothesis of stationarity of the process: by comparing the distributions of observed points to a random distribution, it seems easy to detect the interactions of repellent or attractions that cause dispersion or spatial concentration phenomena.

However, any conclusions should not be too hasty as it should be kept in mind that the same aggregated or dispersed structures can be obtained with an inhomogeneous Poisson process in which the intensity of the process varies in space but the points are independent of each other (see figure 4.5). A single observation of a spatial distribution does not allow for any distinction between first- and second-order properties of a process in the absence of additional information such as that provided by a model that links a covariate to the intensity. Ellison et al. 1997, showed that natural advantages (involving greater intensity) have an effect on the location of establishments that is indistinguishable from that of positive externalities (causing the aggregation): confusion between these two properties may also concern processes.

One final warning concerns homogeneity. Indeed, initially, the methods developed in spatial statistics consisted of testing for the existence of aggregation or repulsion, assuming the homogeneity of the process: the aim was, therefore, to test a spatial distribution against the null hypothesis of complete spatial randomness (CSR). To analyse such datasets, measurements such as the original  $K$  function, proposed by B.D. Ripley (widely used in statistical literature) are adequate. However, if the null hypothesis of a completely random point distribution is considered too strong, other functions must be favoured. This is the case, for example, for earthquake studies (Veen et al. 2006). Figure 4.7 illustrates 5,970 earthquake epicentres in Iran between 1976 and 2016 (of a magnitude greater than 4.5). This data comes from package *etas*.

---

```
data(iran.quakes, package = "ETAS")
plot(iran.quakes$lat~iran.quakes$long , xlab="Longitude", ylab="Latitude")
```

---

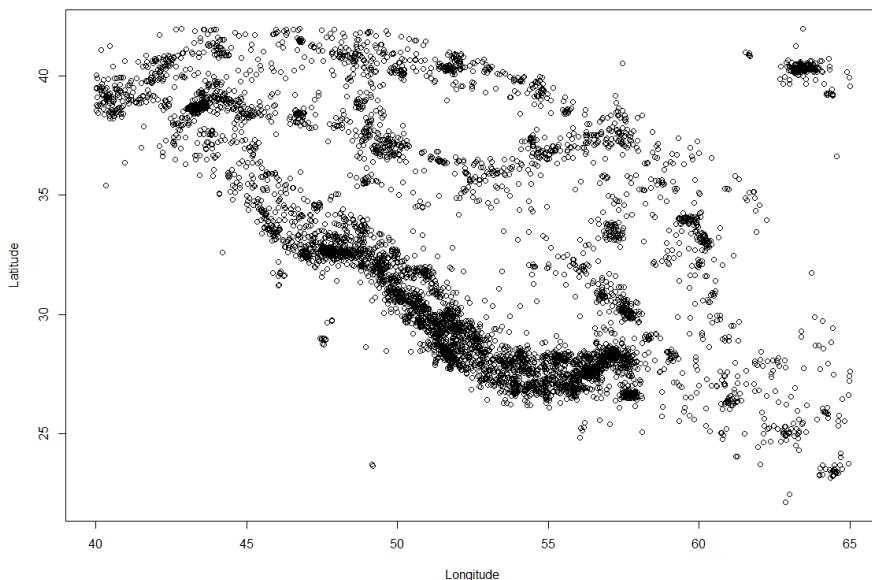


Figure 4.7 – Location of 5,970 earthquake epicentres in Iran from 1976 to 2016

**Source:** package *etas*, authors' calculations.

It is easy to see that any reference to the homogeneity of space is not optimal because there are geological predispositions in this case. B.D. Ripley's  $K$  function would be unsuitable for analysis of this type of data and other tools should be used, such as the *inhomogeneous K* function from Baddeley et al. 2000, which we will present in this chapter. Duranton et al. 2005 also highlighted this limitation of homogeneity of space to analyse the distribution of industrial activities and proposed a new function  $K_d$ .

Thorough knowledge of the available functions is, therefore, essential to characterise point distribution *accurately*. This will be the subject of the next section.

## 4.4 What statistical tools should be used to study spatial distributions?

Unfortunately, the answer to this question is not straightforward. The answer lies in precise analysis of the question that we are attempting to answer, using distance-based measures (particularly with regards to the reference value) and examination the properties of the functions. To fully understand this point and, therefore, the difficulty associated with the choice of the measure, this section will begin with a presentation of the original Ripley's  $K$  function and significant developments that have resulted from this work (sections 4.4.1 and 4.4.2). We will then take time to better explain the determining factors in the choice of measure (section 4.4.3). We will then see the advantages and disadvantages of the existing measures. For an overview of the literature or an in-depth and more complete comparison of measurements, please refer to the work of Baddeley et al. 2015b or the typology of distance-based measures proposed by Marcon et al. 2017.

### 4.4.1 Ripley's $K$ function and its variants

The most widely used indicator for illustrating correlation in point processes is the  $\hat{K}$  empirical function, proposed by B.D. Ripley in 1976 (Ripley 1976; Ripley 1977). This function is commonly known as **Ripley's function** and has been the subject of many comments and developments and several variants. Specifically, this function will allow us to estimate the average number of neighbours relative to the intensity.

**Definition 4.4.1 — Ripley's K function.** Its estimator is written as follows:

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_i \sum_{j \neq i} \mathbf{1}\{\|x_i - x_j\| \leq r\} c(x_i, x_j; r), \quad (4.10)$$

where  $n$  is the total number of points in the observation window,  $\mathbf{1}\{\|x_i - x_j\| \leq r\}$  is an indicator that is worth 1 if points  $i$  and  $j$  are at least equal to  $r$  and 0 otherwise.  $c(x_i, x_j; r)$  corresponds to the correction of edge effects and  $W$  to the study area.

$K$  is a **cumulative function**, giving the average number of points at a distance less than or equal to  $r$  from any point, **standardized by the intensity of the process ( $n/|W|$ )**, which is **assumed to be homogeneous**.

In practical terms, to study the neighbourhood of points, we will analyse all the distances  $r$ , by calculating the value of the  $K$  function for each of these distances. This is done as follows:

1. for each point and distance  $r$ , the number of neighbours (other points) located on the circle with radius  $r$  is counted;
2. we then calculate the *average* number of neighbours (taking into account any edge effects) for each distance  $r$ ;
3. lastly, these results will be compared to those obtained on the assumption of a homogeneous distribution (completion of a homogeneous Poisson process), which will be the expected reference value.

Finally, we will try to detect if there is a significant difference between the estimates of the observed and expected number of neighbours.

In Figure 4.8 we compare the three typical spatial distributions that we considered previously and the three resulting  $K$  function curves. The distance  $r$  is represented graphically in abscissa and the value of the  $K$  function estimated at this distance is represented in ordinate. With package *spatstat*, the  $K$  function is calculated using the *Kest.* function. In Figure 4.8, the estimated  $K$  function is shown in black on the three graphs and the reference value in red dotted lines.

Findings:

- **when the process is completely random, the curve deviates relatively little from  $\pi r^2$ .** This can be seen on the graph at the bottom left of Figure 4.8. The  $K$  curve remains close to the reference value  $\pi r^2$ , for all radii  $r$ .
- **in the case of a regular process,** we obtain:  $\hat{K}(r) < K_{pois}(r)$  because if the points are repulsive, they have fewer neighbours on average in a radius  $r$  than they would have based on the assumption of a random distribution of points. Graphically, the  $K$  curve reflects this repulsion: we see on the right-hand graph that the  $K$  curve is located below the reference value ( $\pi r^2$ ) for all radii.
- **in the case of an aggregated process,** there are on average more points in a radius  $r$  around the points than the expected number under a random distribution: consequently, the points attract each other and  $\hat{K}(r) > K_{pois}(r)$ . Graphically, the  $K$  curve is this time located above the reference value for all areas of study, as can be seen on the central graph shown in Figure 4.8.

---

```
library("spatstat")
par(mfrow=c(2, 3), mar=c(1, 2, 2, 2))
plot(rpoispp(50), main="")
plot(rMatClust(5, 0.05, 10), main="")
```

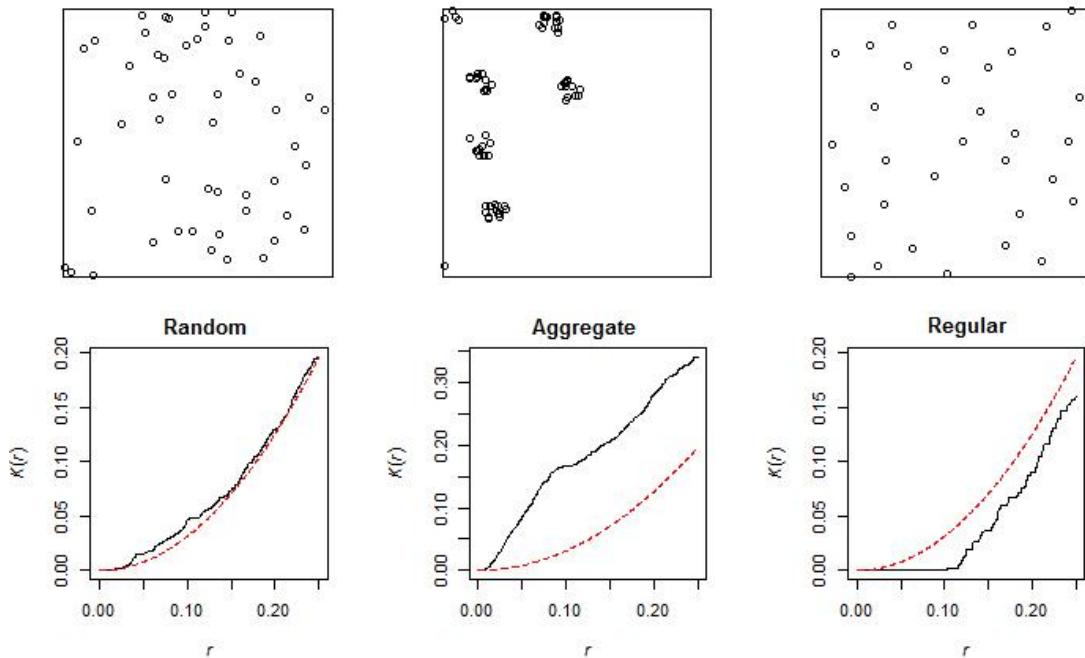


Figure 4.8 –  $K$  functions for the three standard configurations of points

**Source:** package *spatstat*, authors' calculations

```
plot(rMaternII(200,0.1), main="")
par(mar=c(4, 4.1, 2, 3))
# Function K calculated by spatstat
plot(Kest(rpoispp(50),correction="isotropic"),legend=FALSE,main="Random")
plot(Kest(rMatClust(5, 0.05, 10),correction="isotropic"),legend=FALSE,main=
  "Aggregate")
plot(Kest(rMaternII(200,0.1),correction="isotropic"),legend=FALSE,main=
  "Regular")
par(mfrow=c(1, 1))
```

Let's go over a few important points.

Firstly, the  $K$  function is defined using the (strong) stationarity hypothesis. In the case of an inhomogeneous Poisson processes, the difference from the empirical function may be due to the variation in intensity rather than to a phenomenon of attraction, *i.e.* related to the second order property.

Similarly, the interpretation is subject to the same questions as for “conventional” statistics. Correlation does not lead to causality. A lack of correlation does not necessarily lead to independence either.

In addition, the cumulative nature of the function  $K$  must be taken into account. A high  $K$  value at distance  $r_0$  may be due to the combination of phenomena at smaller distances, whereas no interaction exists between points far from  $r_0$ .

Note that there is a **link between the  $K$  function and the point pair correlation function**. This can be approached as follows: draw two concentric circles with radii  $r$  and  $r + h$ , and you count the points in the resulting ring. The expected number is  $\lambda K(r + h) - \lambda K(r)$  If the expression

is standardised by the expected value in the ring for a Poisson process, we obtain:

$$g_h(r) = \frac{\lambda K(r+h) - \lambda K(r)}{\lambda \pi(r+h)^2 - \lambda \pi r^2} = \frac{K(r+h) - K(r)}{2\pi rh + \pi h^2}. \quad (4.11)$$

If we make  $h$  tend to 0,  $g(r) = \frac{K'(r)}{2\pi r}$  or  $K(r) = \int_0^1 sg(s)ds$ , the link between the  $g$  function and the  $K$  function is clear.

Finally, the values returned by the  $K$  function enable possible interactions to be detected between the points for each of the distances studied, on *the whole* of the analysed territory. However, it may be worthwhile to have local information, as for surface data models for which we calculate local indicators known as LISA alongside spatial autocorrelation indicators (such as Moran) (see Chapter 3: "Spatial autocorrelation indices"). In point models, **there are also local indicators built on the principle of Ripley indicators**. An indicator is calculated for each point  $\hat{K}(r, x_i)$ . The only pairs of points taken into account are those that contain the point  $x_i$ . One of the local values or all the values can then be represented graphically. Different points can be identified graphically or even by using functional data analysis methods.

### The $L$ function of Besag 1977

The particular interest of the Ripley function and more generally of distance-based methods lies in the fact that they analyse the space studied by running *all distances* and not using just one or a few geographical levels. The spread of points is very carefully studied and no analysis distance is omitted. Consequently, **only these methods can be used to detect exactly at what distance(s) attraction or dispersion phenomena are observable, with no scale bias associated with a predefined zone**. If there are, for example, aggregates of aggregates in spatialized data, such functions can detect the distances at which spatial concentrations occur: down to the size of the aggregate and the distance between aggregates. More complex spatial structures may also be detected, such as multiple agglomeration phenomena for certain distances and repulsion for other distances (this will be the case if several aggregates are regularly spaced, for example). An additional benefit is to be able to compare the values produced by the functions between several distances. This can be done with the  $K$  function. In the original version of the  $K$  function, it is not easy to directly compare the estimated values for several areas because the reference value,  $\pi r^2$ , requires new calculations (since hyperbolic graphic comparisons are not immediate). As we will see, this has been one of the motivations for development of Ripley's original function.

Two transformations of the Ripley function are frequently used. It is not uncommon to find applications with these variants in statistical literature rather than the original  $K$  function (*e.g.* Arbia 1989 concerning the distribution of industrial companies, Goreaud et al. 1999 concerning the distribution of trees or Fehmi et al. 2001 for plants). The first variant is the  $L(r)$  function proposed by Besag (Besag 1977), which is defined by:  $L(r) = \sqrt{\frac{K(r)}{\pi}}$ , and which is valid in a random process  $L_{Pois}(r) = r$ . With package *spatstat*, the  $L$  function can be calculated using the *Lest* function. Another possible version is  $L(r) - r$ , which is compared to 0 in the case of a completely random distribution. The two advantages to these variants are one the one hand a more stable variance (Goreaud 2000) and, on the other hand, almost immediately interpretable results (Marcon et al. 2003). For example, by using the second variant, if the  $L(r) - r$  function reaches 2 for a radius  $r$  of 1, this means that on average there are as many neighbours within a radius of 1 around each point in this configuration as there would be in a radius of 3 (=2+1) if the distribution were homogeneous. A better standardisation is  $\frac{K(r)}{\pi r^2}$  whose expected value is 1 and whereby the empirical value is the ratio between the number of neighbours observed and neighbours expected (Marcon et al. 2017).

By way of illustration, we have used the example of an aggregated distribution that was given in Figure 4.9, with the four estimated results of the functions  $K$ ,  $L$ ,  $L - r$  and  $K(r)/\pi r^2$  for this distribution.

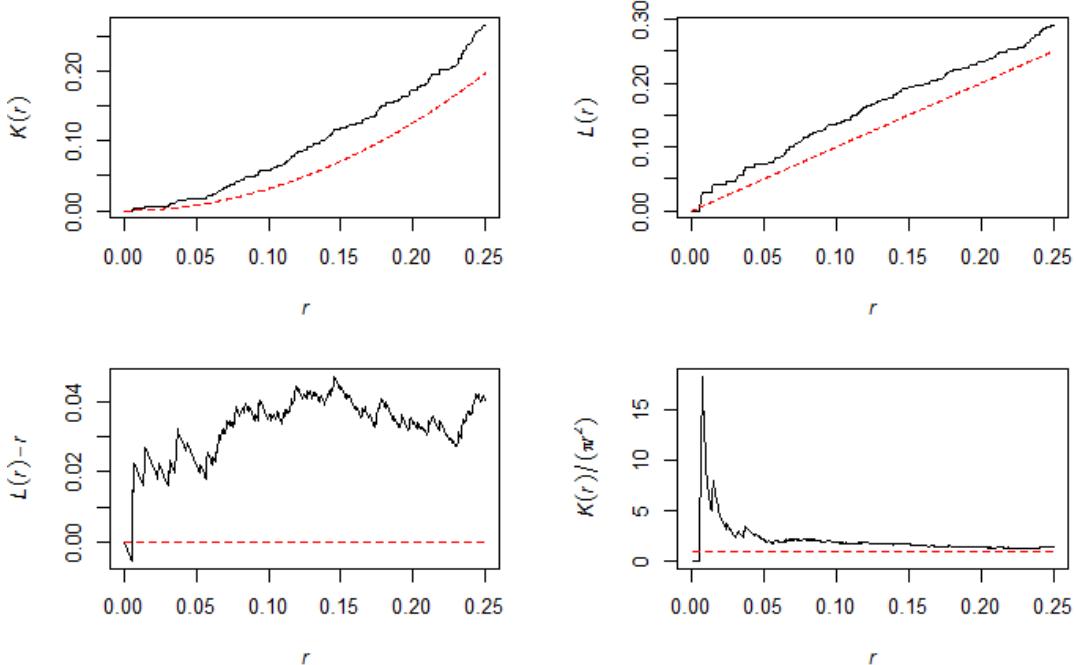


Figure 4.9 – Representation of functions  $K$ ,  $L$ ,  $L - r$  and  $K(r)/\pi r^2$

**Source:** package *spatstat*, authors' calculations

---

```
library("spatstat")
AGRE<- rMatClust(10, 0.08, 4)
K<- Kest(AGRE, correction="isotropic")
L<- Lest(AGRE, correction="isotropic")
par(mfrow=c(2, 2))
plot(K, legend =FALSE, main="") # K
plot(L, legend =FALSE, main="") # L
plot(L, .-r~ r, legend =FALSE, main="") # L defined as L(r)-r
plot(K, ./ (pi*r^2)~ r, legend =FALSE, main="") # K(r)/(pi r^2)
par(mfrow=c(1, 1))
```

---

### The $D$ function of Diggle et al. 1991

$K$  and  $L$  functions may be used in the studies if the hypothesis of homogeneity of the analysed space is verified. Another variant of the  $K$  function allows the non-homogeneity of space to be taken into account: this is the  $D$  function as proposed by Diggle et al. 1991. This indicator is directly derived from the work of epidemiologists, seeking to compare the concentration of “cases” (children with a rare disease in North Britain) and “controls” (healthy children in the same study area). This function is very simply defined as the difference between two Ripley  $K$  functions: cases and controls. We obtain:

$$D(r) = K_{cas}(r) - K_{controls}(r) \quad (4.12)$$

The  $D$  function enables distributions of two sub-populations to be compared. Intuitively, it is understood that if cases are more localised than controls, a spatial concentration of cases will be

detected by the  $D$  function. Conversely, if the distribution of cases is less concentrated than that of controls, the  $D$  function will detect that cases will be spatially more dispersed than controls. The benefit of using this function is to be able to detect differences in the distribution being studied compared to a reference distribution. This may be interesting, for example, if we want to know whether a certain type of housing is geographically more concentrated than other types of housing, or whether a type of business is more concentrated within cities than other types of businesses etc. The difference in two  $K$  functions gives a comparison value for  $D$  of 0 for all areas of study. However, it is impossible to compare the estimated  $D$  values due to changes in the reference sub-population. This  $D$  function can be implemented in the R software using package *dbmss*: we will then use the function called *Dhat*. Just like the  $K$  function, it is also possible to associate a level of significance of the results by randomly labelling points (see below). The *DEnvelope* function will then be favoured. Various applications are available in the literature regarding the spatial concentration of economic activities (such as Sweeney et al. 1998). Interested readers may also find a variant of the  $D$  function proposed by Arbia et al. 2008.

#### The $K_{inhom}$ function of Baddeley et al. 2000.

$K_{inhom}$ , the version of Ripley's  $K$  function in inhomogeneous space was proposed by Baddeley et al. 2000. The estimated value of  $K_{inhom}$  therefore involves the estimated values of the intensity (the hypothesis of an identical intensity at any point in the territory being studied must be rejected since the space in question is no longer homogeneous). By noting  $\widehat{\lambda}(x_i)$  as the estimation of the process around point  $i$  and  $\widehat{\lambda}(x_j)$  as the estimation of the process around point  $j$ , the cumulative function  $K_{inhom}$  can be defined as follows:

$$\widehat{K}_{inhom}(r) = \frac{1}{D} \sum_i \sum_{j \neq i} \frac{\mathbf{1}\{\|x_i - x_j\| \leq r\}}{\widehat{\lambda}(x_i)\widehat{\lambda}(x_j)} e(x_i, x_j; r) \quad (4.13)$$

$$\text{with } D = \frac{1}{|W|} \sum_i \frac{1}{\widehat{\lambda}(x_i)}.$$

We can show that in the case of an inhomogeneous process:  $K_{inhom,pois}(r) = \pi r^2$ . Estimates of  $K_{inhom}$  are therefore interpreted in the same way as in the case of the homogeneous  $K$  function. From a practical point of view, the *Kinhom* function in package *spatstat* enables the  $K_{inhom}$  function to be calculated.

In theory, the treatment of non-stationary processes could be considered resolved, but the practical difficulty lies in estimating local densities using the kernel method. Beyond the technical difficulties, the theoretical impossibility of separation in a single observation based on first order phenomenon (intensity) and based on aggregation of the phenomenon being studied results in significant biases when the window used to estimate local densities is of the same order of magnitude as the  $r$  value in question. There are still few empirical applications for this indicator (Bonneu 2007; Arbia et al. 2012).

#### 4.4.2 How can we test the significance of the results?

Several statistical methods can be used to assess the significance of the results obtained using the various, previously presented functions. The most common technique is the use of the Monte Carlo method to simulate a confidence interval, which we will begin by explaining.

##### Monte Carlo methods

Without any knowledge of the theoretical distribution of Ripley's  $K$  function under the null hypothesis of a completely random distribution, **the significance of the difference between observed values and theoretical values is tested by the Monte Carlo method**. This method can

be used to determine confidence intervals for all derivative functions of  $K$  that have been presented. The function in question will be designated generically by  $S$ . To do this, we proceed as follows:

1. A number  $q$  of datasets is generated, corresponding to the null hypothesis of the test. If the null hypothesis is a completely random process, we generate  $q$  Poisson intensity processes, corresponding to the spatial distribution being tested.
2. Curves  $U(r) = \max\{S^{(1)}(r), \dots, S^{(q)}(r)\}$  and  $L(r) = \min\{S^{(1)}(r), \dots, S^{(q)}(r)\}$  are defined, which can be used to define an envelope, represented in grey in the graphs produced with the R software.
3. For a bilateral test, the defined envelope corresponds to a first type of risk  $\alpha = \frac{2}{q+1}$ , i.e. 39 simulations for a test at 5 %.

For each of the functions, we can build this envelope that allows us to compare the statistics built from the data to statistics derived from the simulation of a random process corresponding to the tested null hypothesis (a homogeneous Poisson process of the same intensity for the function  $K$ ). In package *spatstat*, the generic *envelope* command is used to run Monte Carlo simulations and construct curves corresponding to the upper and lower values of the envelope. The envelope should not be interpreted as a confidence interval around the indicator being studied: it indicates the critical values of the test. To give a simple example, let's use dataset *paracou16* relating to the location of trees in the Paracou forest research station in French Guiana. This data is available in package *dbmss*. Let's calculate the confidence interval associated with the  $K$  function with 39 simulations. In Figure 4.10, the obtained  $K$  curve is shown (full black line), the red dotted curve represents the middle of the confidence interval and the two envelope markers are given as well as the envelope (curves and grey envelope). We can see that, up to a distance of close to 2 metres, we cannot reject the null hypothesis of a CSR process based on the Ripley function.

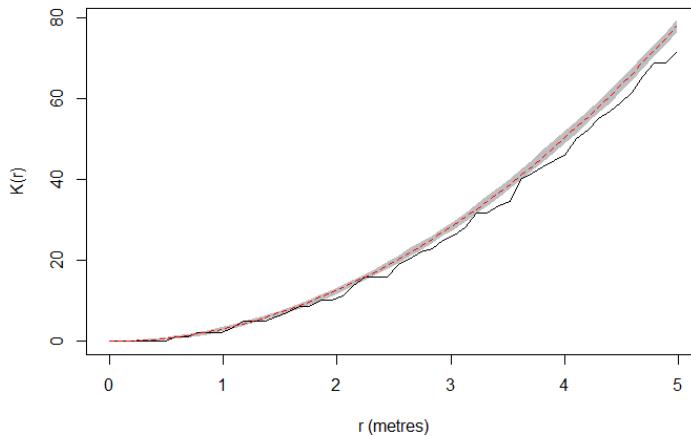


Figure 4.10 – Example of a confidence envelope for the  $K$  function

**Source:** packages *spatstat* and *dbmss*, *paracou16* data, authors' calculations

---

```
library("dbmss")
# Envelope calculated using package dbmss, data: 2,426 points.
env <- KEnvelope(paracou16, NumberOfSimulations=39)
plot(env,legend =FALSE, main="", xlim=c(0,5), xlab = "r (metres)", ylab= "K
(r)")
```

---

With increased computing power, a common practice is to simulate the null hypothesis many times (1,000 or 10,000 times rather than 39) and to define the envelope from quantiles  $\alpha/2$  and  $1 - \alpha/2$  for values of  $S(r)$ .

The test is repeated for each value of  $r$ : the risk of mistakenly rejecting the null hypothesis is therefore increased beyond  $\alpha$ . This underestimation of the first type of risk is not very large because the values of the cumulative functions are auto-correlated to a great degree. The test is therefore commonly used without any particular precautions. However, authors such as Duranton et al. 2005 consider this to be serious and try to remedy it. A method to correct the problem is presented in Marcon et al. 2010 and implemented in package *dbmss* under the name of the overall confidence interval of the null hypothesis (as opposed to the local confidence intervals calculated at each  $r$  value). It consists of repeatedly removing a part  $\alpha$  of simulations of which at least one value contributes to  $U(r)$  or  $L(r)$ .

One important point: when calculating an envelope under R, it is systematically associated with a particular function. In other words, the calculation routines available in the packages take into account the specific nature of the functions: the confidence intervals are therefore simulated by considering the correct null hypothesis. For example, to simulate the envelope for the  $K$  function, the null hypothesis is constructed from points that are distributed randomly and independently in the study area. However, for the  $D$  function of Diggle et al. 1991, to develop a confidence interval with the same assumptions as for the  $K$  function would be incorrect. For  $D$ , you must take into account variations in intensity in the area studied. What next? Remember that the null hypothesis for this function corresponds to a situation where the sub-population of the cases and the sub-population of the controls have the same spatial distribution. The solution suggested by Diggle et al. 1991 is random labelling which involves, for each simulation, assigning a “case” or “control” label for each location. This random permutation of labels in unchanged locations is a quite intuitive technique that will also be used to develop confidence intervals for other functions that we will study in section 4.5. Under R, packages *spatstat* or *dbmss* have options for calculating functions that allow this hypothesis of random labelling to be simulated.

### Analytical tests

There are few analytical tests in the literature and they are rarely used in studies, even though they have the advantage of saving calculation time for confidence intervals. For  $K$ , for example, analytical tests exist in simple areas of study (Heinrich 1991). In the particular case of the CSR character test in a rectangular window, Gabriel Lang and Eric Marcon recently developed a classic statistical test (Lang et al. 2013) available in the *Ktest* function of package *dbmss* (Marcon et al. 2015b). It returns the probability of mistakenly rejecting the null hypothesis of a completely random distribution from a spatial distribution, without using simulations: the distribution of the  $K$  function with no correction for side effects follows an asymptotically normal distribution of known variance. The test can be used from a few dozen points. It should be noted that such tests for lesser known functions are also proposed in the literature (Jensen et al. 2011).

#### 4.4.3 Review and focus on important properties for new measurements

Measures derived from the Ripley’s  $K$  function are useful in many configurations to explain the interactions between the points studied. We have, incidentally, given many references in various areas of application. However, specific developments can still be considered to answer certain questions, such as the location of economic activities. To understand this point, we will consider the strengths and limitations of the measures taken by Ripley’s  $K$  function as part of this framework of analysis.

**Review: Are the functions derived from Ripley's  $K$  suitable for describing the spatial concentration of economic activities?**

The statistical tools presented in the previous sections are valuable, but their use in illustrating data for equipment or companies is not straightforward. To further consider this notion, let's go back to the examples in the introduction (the four facilities) and select Ripley's  $K$  function to characterise the spatial structures of each of these facilities. The results are shown in Figure 4.11: the estimated function of Ripley's  $K$  is shown as a continuous line, the confidence intervals obtained from 99 simulations by the grey area, the centre of the confidence interval is indicated by the dotted curve and the edge effects were calculated by the Ripley method. This correction of edge effects is based on the idea that, for a given point, the part of the crown outside the area (see Figure 4.2) contains the same density of neighbours as the part located within the study area. This hypothesis is acceptable because, let's remember, we consider there to be a completely random point distribution in the case of Ripley  $K$  function.<sup>5</sup>

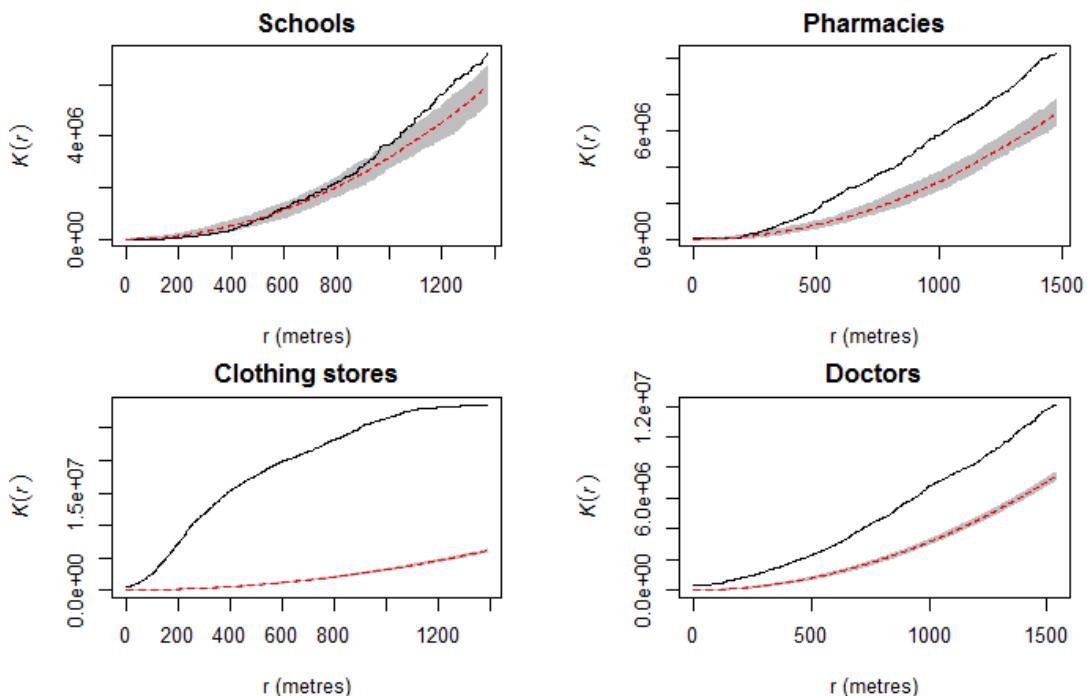


Figure 4.11 – Ripley functions for the four facilities

**Source:** INSEE-BPE, packages *spatstat* and *dmbs*, authors' calculations

```
library("dmbs")
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))
bpe_sch<- bpe[bpe $TYPEQU=="C104", ]
bpe_phar<- bpe[bpe $TYPEQU=="D301", ]
bpe_clo<- bpe[bpe $TYPEQU=="B302", ]
bpe_doc<- bpe[bpe $TYPEQU=="D201", ]

schools <- as.ppp(bpe_sch[,c ("lambert_x", "lambert_y")], owin(c(min(bpe_sch[, "lambert_x"]), max(bpe_sch[, "lambert_x"])), c (min(bpe_sch[, "lambert_y"]), max(bpe_sch[, "lambert_y"])))
```

5. Technically, let us assume that a neighbour of a given point is located in the crown width (inside the domain). The Ripley correction consists in assigning to this neighbour a weight equal to the inverse of the ratio of the perimeter of the crown over the total perimeter of the crown.

```

lambert_y"])), max(bpe_sch[, "lambert_y"])))
bpe_schools_wmppp <- as.wmppp(schools)
pharma <- as.ppp(bpe_pharm[, c("lambert_x", "lambert_y")], owin(c(min(bpe_
    pharm[, "lambert_x"]), max(bpe_pharm[, "lambert_x"])), c(min(bpe_pharm[, "_
        lambert_y"]), max(bpe_pharm[, "lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
clothing <- as.ppp(bpe_clo[, c("lambert_x", "lambert_y")], owin(c(min(bpe_
    clo[, "lambert_x"]), max(bpe_clo[, "lambert_x"])), c(min(bpe_clo[, "_
        lambert_y"]), max(bpe_clo[, "lambert_y"]))))
bpe_clothing_wmppp <- as.wmppp(clothing)
doctors <- as.ppp(bpe_doc[, c("lambert_x", "lambert_y")], owin(c(min(bpe_
    doc[, "lambert_x"]), max(bpe_doc[, "lambert_x"])), c(min(bpe_doc[, "_
        lambert_y"]), max(bpe_doc[, "lambert_y"]))))
bpe_doctors_wmppp <- as.wmppp(doctors)

kenv_schools <- KEnvelope(bpe_schools_wmppp, NumberOfSimulations=99)
kenv_pharma <- KEnvelope(bpe_pharma_wmppp, NumberOfSimulations=99)
kenv_clothing <- KEnvelope(bpe_clothing_wmppp, NumberOfSimulations=99)
kenv_doctors <- KEnvelope(bpe_doctors_wmppp, NumberOfSimulations=99)
par(mfrow=c(2, 2))

plot(kenv_schools, legend=FALSE, main="Schools", xlab = "r (metres)")
plot(kenv_pharma, legend=FALSE, main="Pharmacies", xlab = "r (metres)")
plot(kenv_clothing, legend=FALSE, main="Clothing stores", xlab = "r (metres
    )")
plot(kenv_doctors, legend=FALSE, main="Doctors", xlab = "r (metres)")
par(mfrow=c(1, 1))

```

The results obtained in Figure 4.11 confirm the notions that we had regarding the spatial distribution of each of the facilities in Rennes (see Figure 4.1). For doctors, clothing shops and pharmacies, significant levels of spatial concentration are detected (graphically, the  $K$  curves are located above the confidence interval). With regard to schools, the trend towards concentration as well as dispersion is not evident since the  $K$  curve for this sector remains within the confidence interval below a radius of one kilometre and then, beyond this radius, the observed distribution of schools in Rennes does not seem to deviate significantly from a random distribution. Finally, note that the spatial concentration is particularly high for clothing stores (the difference between the  $K$  curve and the upper band of the confidence interval is greatest in this sector).

**Can we consider these results sufficient to describe the spatial structure of these facilities or should they be pursued further?** The answer is simple: these conclusions are based on statistically correct calculations, but they may seem economically irrelevant. These results come up against several important limits, in particular the hypothesis of homogeneity. First of all, remember that a spatial concentration detected with the Ripley  $K$  function satisfies a particular definition here: the distributions observed are more concentrated than they would be under the hypothesis of random distribution. This null hypothesis may seem very strong. Let us consider the case of the location of pharmacies: we know that in France that this has to meet certain regulatory provisions linked to the population. The CSR reference distribution does not, therefore, appear to be the most relevant in this case. A solution would then be to take into account this non-homogeneity of the space, for example by retaining the function  $D$  of Diggle et al. 1991 to compare the distribution of pharmacies with that of residents. Provided that the data are available and accessible, this

would allow us to monitor the heterogeneity of the territory. This technique would also make it possible for us to regulate to a certain extent the severe constraints of setting up operations (which would prevent an equal probability of being located at any point in the territory analysed) such as the impossibility of being located in non-buildable areas in Rennes, in urban parks, etc.: the population and shops cannot be located there. It has to be said that although this strategy is attractive, it is not completely satisfactory. For example, in the case of facilities, and even more so in the case of companies, we have observations that are generally weighted very differently (number of employees, etc.). It is therefore difficult to consider that the points analysed all have the same characteristics. However, all the functions presented to date ( $K$ ,  $L$ ,  $D$  and  $K_{inhom}$ ) cannot include a weighting of points. This observation may be very problematic, especially considering that studies of industrial concentration in the sense of Ellison et al. 1997, Maurel et al. 1999 brought together economists' and spatial statisticians' concerns in the late 1990's toward zoning-based spatial concentration indicators. Further developments in this regard must therefore be made for the measures resulting from Ripley's  $K$ .

#### **Development of distance-based measures to meet economic criteria**

In the 2000s', **lists of economically relevant criteria** were proposed to characterise the spatial concentration of economic activities (Duranton et al. 2005; Combes et al. 2004; Bonneu et al. 2015) as:

- the insensitivity of the measure to a change in the definition of geographical scales;
- the insensitivity to a change in definition of sectoral level (according to the selected sectoral classification);
- the comparability of results between sectors;
- taking into account the productive structure of industries (*i.e.* industrial concentration in the sense of Ellison et al. 1997 which depends on both the number of establishments within the sectors and the workforce);
- a reference must be clearly established.

These questions have been discussed in many studies, in particular to distinguish between **appreciable criteria** such as the comparability of results between sectors, **essential criteria** such as the criterion regarding insensitivity of the measure following a change in the definition of geographical scales (this refers to the previously mentioned MAUP). The benefit of all distance-based measures presented in this chapter is avoiding the pitfall of MAUP. On the other hand, no measure has yet tackled sectoral divisions: the problem raised by the second criterion in the above list therefore remains intact.

#### **What research options exist for extending the presented measures?**

Several significant developments were proposed in the 2000s. Continued work by spatial statistical specialists and the inclusion of spatial concerns in economic studies have contributed to important innovations in concentration indicators. Not all of the studies will be covered in this context, but we will consider some of the most widely used information. In the first instance, we will introduce a slightly counter-intuitive notion of the **reference value**. When we try to characterise a point distribution, we implicitly compare it with a reference distribution (the statistician's null hypothesis) and it is the difference from this theoretical distribution that makes it possible to assess the geographical concentration, the dispersion or if the difference is not sufficient to conclude if there is any interdependence between the points. Let's re-examine the example of clothing stores and look at three types of indicators (Marcon et al. 2015a; Marcon et al. 2017) to characterise their location:

- the **topographical measures** use physical space as a reference value (Brülhart et al. 2005). The number of neighbours of points of interest is relative to the surface area of the neighbourhood in question: this is part of the mathematical framework of point processes. This kind of analysis allows the following question to be answered: is the density of clothing shops high around footwear stores? A positive response, for example, will show a topographical concentration of clothing stores (in the vicinity of these stores, the density of clothing stores is high). The measures presented in  $K$ ,  $L$ ,  $D$  and  $K_{inhom}$  accommodate this topographical definition of the reference value (depending on the functions, the theoretical density is considered to be constant or not). It is interesting to note that, for this reference value, the hypothesis of a homogenous or inhomogeneous space can be used;
- the **relative measures** use a distribution that is not physical space as a reference value. The number of neighbours is not shown on the surface, but in the number of points in the reference distribution. This is a clear departure from the theory of point processes, except to consider the reference distribution as an estimate of the intensity of the process based on the null hypothesis of independence between the points. In our example, this amounts to testing the existence of an over-representation or under-representation of clothing stores in the vicinity of clothing stores compared to a reference, such as all business activities. Note: the  $D$  function is not a relative measure under these hypotheses as it compares density to another density, based on difference. On the other hand, a relative measure would answer the following question: around clothing stores, is the frequency of clothing stores greater than average, throughout the territory? A positive response leads us to conclude that there is a relative concentration of clothing stores;
- lastly, **absolute measures** do not require any standardisation (by space or by comparison with any other reference). In our example, this amounts to simply counting the number of clothing shops around the clothing stores. The number obtained can then be compared to its value under the chosen null hypothesis, obtained using the Monte Carlo method.

Based on the works presented above, in particular regarding the  $K$  function, statistical indicators have been proposed in the statistical literature to characterise these spatial structures under the three reference values, as mentioned above (Marcon et al. 2017). We will develop several indicators in the following sections and we will see that another important difference lies in the notion of neighbourhood. For example, it is possible to study the proximity of the points analysed *up to* a certain distance  $r$ . In practical terms, this means characterising the proximity of points on discs of radius  $r$ , which defines cumulative-type functions (such as Ripley's  $K$ ). Another possibility is to assess the proximity of the points not *up to* a distance  $r$  but *at* a certain distance  $r$ . Neighbourhood is assessed in a crown (also called a ring) and density functions are used to characterise it (like the  $g$  function that we have already considered). A graphic illustration of these two definitions is given in Figure 4.12. In the figure on the left, the grey area corresponds to the surface of a disc with radius  $r$  and, in the figure on the right, to the surface of a crown with a radius  $r$ .

The choice of neighbourhood is not insignificant. Therefore, density functions are more precise around the study radius but do not provide information on spatial structures at smaller distances, unlike cumulative functions. Only a cumulative function may, for example, detect whether aggregates are randomly located or whether there is a spatial interaction between aggregates (*e.g.* aggregates of aggregates). However, as cumulative functions accumulate spatial information up to a certain distance, local information at the radius  $r$  is unclear, unlike density functions. The use of one or other of these neighbourhood concepts has advantages and disadvantages (Wiegand et al. 2004; Condit et al. 2000).

Marcon et al. 2017 proposed an initial classification of distance-based functions according to these two criteria:

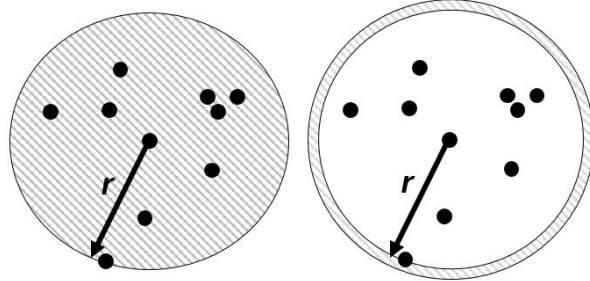


Figure 4.12 – Two possible neighbourhood concepts: on a disk or on a crown

**Source:** the authors

- **the type of function:** probability density, like the  $g$  function or cumulative, such as Ripley's  $K$  function;
- **the reference value** that can be topographical (Ripley functions and their direct variants), related to a reference situation (such as  $M$  that we will present in the next section) or absolute (*i.e.* without reference such as the  $K_d$  function, also presented in the next section).

It is easy to see why the choice of the correct measure is not immediately clear: first of all, the question being asked must be identified in order to select the most appropriate measure.

## 4.5 Recently proposed distance-based measures

In this section, we will present two measures relating to two references that have not yet been dealt with: the absolute and relative reference.

### 4.5.1 The $K_d$ indicator of Duranton and Overman

Unlike the previously presented functions, this indicator was developed by economists and was drawn up without any direct links with Ripley's work (although it was referred to in the bibliography). The idea of this function is to be able to estimate the probability of finding a neighbour at a distance  $r$  from each point.

**Definition 4.5.1 —  $K_d$  function of Duranton and Overman.** Through standardisation, Duranton et al. 2005 define  $K_d$  as a function of density of probability of finding a neighbour at a distance  $r$ . This function can therefore be qualified as an absolute measurement of density because it has no reference. The proposed indicator is written:

$$K_d(r) = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} \kappa(\|x_i - x_j\|, r) \quad (4.14)$$

with  $n$  designating the total number of points of the sample and  $\kappa$ , the Gaussian kernel as

$$\kappa(\|x_i - x_j\|, r) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{(\|x_i - x_j\| - r)^2}{2h^2}\right).$$

Here we can see the technical difficulty of counting neighbours at a distance  $r$  because it requires the use of a smoothing function (hence the use of the Gaussian kernel in the function). This smoothing function allows you to count neighbours whose distance is "around"  $r$ . The bandwidth can be defined in several ways but the Silverman 1986 method is mentioned in the original article of Duranton and Overman. As with other distance-based functions, a confidence interval of the null hypothesis can be assessed to assess the significance of the results obtained. The marks (weight/type

pairs) are redistributed to all existing locations (positions taken by points): this technique makes it possible to control both the industrial concentration and the general location trends of all types of points (two properties listed in the “correct” concentration index criteria applicable to economic activities). The hypothesis of a random location of type  $S$  points is rejected at distances  $r$ , if the function  $K_d$  is located above or below the trust boundary of the null hypothesis. Another version of  $K_d$  that takes into account the weighting of points exists - it was proposed in the original article by Duranton et al. 2005. Behrens et al. 2015 used a cumulative function  $K_d$ . It should be noted that the  $K_d$  function has been the subject of many empirical applications in spatial economics (e.g. Duranton 2008, Barlet et al. 2008).

The  $K_d$  function can be calculated under R using the `Kdhat` function in package `dbmss`. The `KdEnvelope` function that is available in the same package can be used to associate a confidence interval with the results obtained.

#### 4.5.2 $M$ function of Marcon and Puech

The  $M$  indicator by Marcon et al. 2010 is a cumulative indicator, like Ripley’s  $K$ , as it is calculated by varying a disc of radius  $r$  around each point. This is a relative indicator since it compares the proportion of points of interest in a neighbourhood with the proportion of points seen throughout the territory analysed. If we consider that clothing stores are attracted to each other, their proportion around each clothing store will be higher than in the city. In practice, for a radius  $r$ , we will calculate the ratio between the local proportion of clothing stores around clothing stores and the proportion observed in the city. This calculation is repeated for all clothing stores and the average of these relative proportions is calculated. The reference value for the  $M$  function is 1. A higher value reflects a relative spatial concentration, and a lower value shows a tendency towards repulsion (the minimum value being 0). The values of  $M$  can also be interpreted in terms of ratio comparisons: for example, if  $M(r)=3$ , this indicates that on average there is a 3 times higher frequency of points of interest appearing around points of interest within a radius  $r$  than the frequency observed over the entire observation window. Finally, like the  $K_d$  function,  $M$  can include weighting of points.

**Definition 4.5.2 —  $M$  function of Marcon and Puech.** Formally, for  $S$  type points, Marcon and Puech’s  $M$  function is defined as:

$$M(r) = \sum_{i \in S} \frac{\sum_{j \neq i, j \in S} \mathbf{1}(\|x_i - x_j\| \leq r)}{\sum_{j \neq i} \mathbf{1}(\|x_i - x_j\| \leq r)} / \frac{n_S - 1}{n - 1}. \quad (4.15)$$

where  $n_S$  and  $n$  refer respectively to the total number of  $S$  type points and the total number of all types of points in the study window. This indicator should be read as the result of two frequency reports. The local average of the frequency of  $S$  type points is compared within a radius  $r$  around  $S$  type points with the frequency of  $S$  type points over the entire observation window. Removing a point from the denominator avoids a slight bias, since the centre point cannot always be counted in its neighbourhood.

As for the  $K_d$  function, a version exists that takes into account the weighting of points (Marcon et al. 2017). Technically, this means multiplying the indicator by the weight of the neighbouring point in question (for example, by the number of its employees or its turnover if we look at industrial establishments). As with the other indicators, a confidence interval can be generated using Monte Carlo methods. The specific nature of the points is retained (weight/sector pairing). For  $M$ , as for  $K_d$ , the control for industrial concentration is not present in the definition of the function but in

the definition of the confidence interval, as the points labels (weight/sector pairs) are redistributed to the existing locations. In their latest work, Lang et al. 2015 offered a non-cumulative version of the  $M$  indicator, named  $m$ , similar to the  $g$  function for  $K$ , see Equation 13.8. As in all the situations we have encountered, the indicators can lead to different analyses: since the reference values are not the same, they answer different questions. The analyses provided are, therefore, complementary (Marcon et al. 2015a; Lang et al. 2015). Finally, note that the  $M$  function does not require correction of edge effects and can be calculated in R using the `Mhat` function in package `dbmss`. The `MEnvelope` function in the same package makes it possible to combine a confidence interval to judge the significance of the results obtained.

As an example of application, consider the spatial structures of the four facilities in the introductory example of the city of Rennes. A graphical representation of the results of the  $M$  function for schools, pharmacies, doctors and clothing stores is given in Figure 4.13.

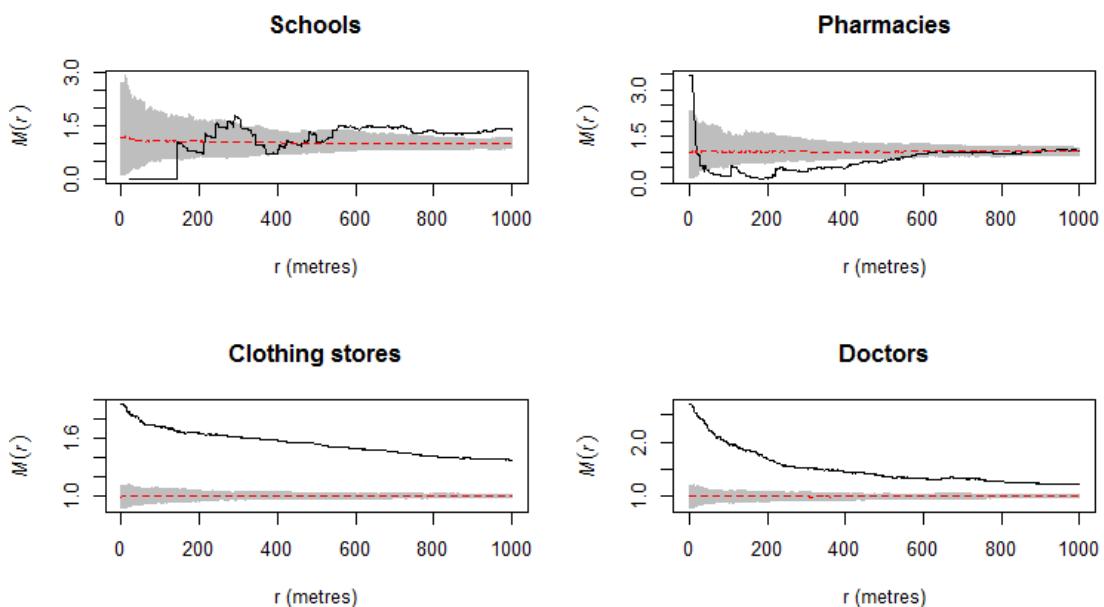


Figure 4.13 – Marcon and Puech functions for the four facilities

**Source:** INSEE-BPE, packages `spatstat` and `dbmss`, authors' calculations

```
library("dbmss")
# Set of marked points
bpe_equip<- bpe[bpe $TYPEQU %in% c ("C104", "D301", "B302", "D201"), c (2,3,1)]
colnames(bpe_equip) <- c("X", "Y", "PointType")
bpe_equip_wmppp <- wmppp(bpe_equip)
r<- 0:1000
NumberOfSimulations<- 99
menv_sch<- MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="C104")
menv_ph<- MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="D301")
menv_clo<- MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="B302")
menv_doc<- MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
```

```

  ReferenceType="D201")
par(mfrow=c(2, 2))
plot(menv_sch, legend=FALSE, main="Schools", xlab = "r (metres)")
plot(menv_phc, legend=FALSE, main="Pharmacies", xlab = "r (metres)")
plot(menv_clo, legend=FALSE, main="Clothing stores", xlab = "r (metres)")
plot(menv_doc, legend=FALSE, main="Doctors", xlab = "r (metres)")
par(mfrow=c(1, 1))

```

It is easy to see that levels of spatial concentration can be seen for all of the distances studied for doctors or clothing stores (both associated  $M$  curves being located above their respective confidence interval up to 1 kilometre). As it is possible to compare the values obtained by the  $M$  function, we can also conclude that the highest levels of aggregation appear at short distances. Thus, in the very first area of study, the proportion of clothing stores around clothing stores is approximately 2 times higher than the proportion of clothing stores observed in the city of Rennes. This result is quite close to the conclusions drawn by Marcon et al. 2015a in Lyon for this activity. With regard to schools or pharmacies, however, concentration or dispersion levels are detected according to the distances in question. Schools, for example, appear dispersed up to approximately 150 metres (the associated  $M$  curve is located beneath the confidence interval of the null hypothesis up to this distance), then, beyond a distance of 500 metres, a phenomenon of spatial concentration is detected. At very short distances, pharmacies appear spatially aggregated, whereas their distribution is dispersed above approximately 50 metres. However, for schools and pharmacies, we note that the  $M$  curves remain fairly close to their respective confidence intervals.

#### 4.5.3 Other developments

This area of statistical literature is currently growing rapidly (Duranton 2008, Marcon et al. 2017). The contributions are varied: statisticians define the necessary theoretical framework and researchers develop tools applicable to their specific field. Among the work carried out recently, Bonneau et al. 2015 propose a family of indicators that have the merit of showing links between the Bonneau-Thomas (proposed in this article), Marcon-Puech and Duranton-Overman indicators. Not all indicators have yet been implemented in the usual software, even if efforts are made to take account of recent developments in the literature and make them freely available to interested users.

### 4.6 Multi-type processes

The introduction presented four maps relating to the respective locations of schools, pharmacies, general practitioners and clothing stores (Figure 4.1, p.73). All this information could have been gathered together, with each activity being a qualitative mark for the process. These marks make it possible to build **multi-type processes**, and to introduce new questions alongside those that have been developed previously: is there independence in location between types (marks)? If the answer is no, are there any phenomena of attraction or repulsion?

In order to provide answers to these questions, we must now consider processes that have specific characteristics: it is therefore possible for us to define indicators of the first order (intensity) and second order (neighbourhood relations), which we will do successively in the following two sub-sections.

#### 4.6.1 Intensity functions

Analysis of variability in the intensity of processes that led to the observation of distribution of the analysed entities is interesting for an initial analysis.

In the field of ecology, one might wonder, for example, (i) if all tree species within a forest are located in the same way, (ii) if the dead trees are more agglomerated than the healthy trees (iii) if

the presence of young shrubs follows that of parent trees etc. The density study gives an initial indication of the observed spatial heterogeneity. In the example below, we have used the respective locations of the trees of a permanent experimental facility in Paracou, French Guiana, available in the Paracou16 dataset in package dbmss. Three tree species are listed: *Vacapoua americana*, *Qualea rosea* and mixed tree species grouped under the term *Other*. The high number of trees present on the Paracou16 plot (2426 trees in total) makes it very difficult to identify any location trends for each species (see Figure 4.14).

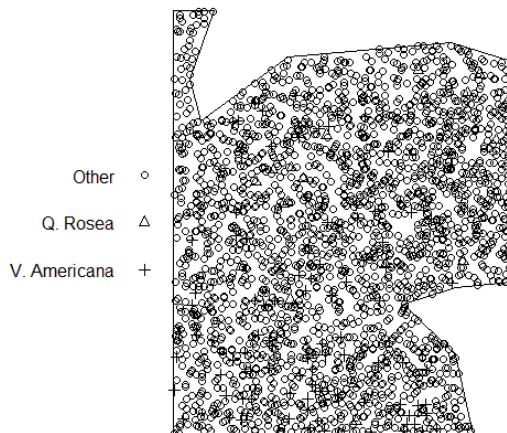


Figure 4.14 – Location of tree species *Vacapoua americana*, *Qualea rosea* or other (mix) in the Paracou16 forest system.

**Source:** Paracou16 data from package dbmss, authors' calculations

---

```
library("dbmss")
data(paracou16)
plot(paracou16, which.marks=2, main = "")
# the 2nd column makes it possible to differentiate the types of points (
  species)
```

---

On the other hand, a representation of the density by species is more informative and makes it possible to highlight differences in location according to the tree species in question (see Figure 4.15). A 2D representation of density is given in this example and obtained from the density function of package spatstat.

---

```
library("dbmss")
data(paracou16)
V.Americana<- paracou16[paracou16$marks$PointType=="V. Americana"]
Q.Rosea<- paracou16[paracou16$marks$PointType=="Q. Rosea"]
Other<- paracou16[paracou16$marks$PointType=="Other"]
par(mfrow=c(1,3))
plot(density(V.Americana, 8), main="V. Americana")
plot(density(Q.Rosea, 8), main="Q. Rosea")
plot(density(Other, 8), main="Other")
par(mfrow=c(1,1))
```

---

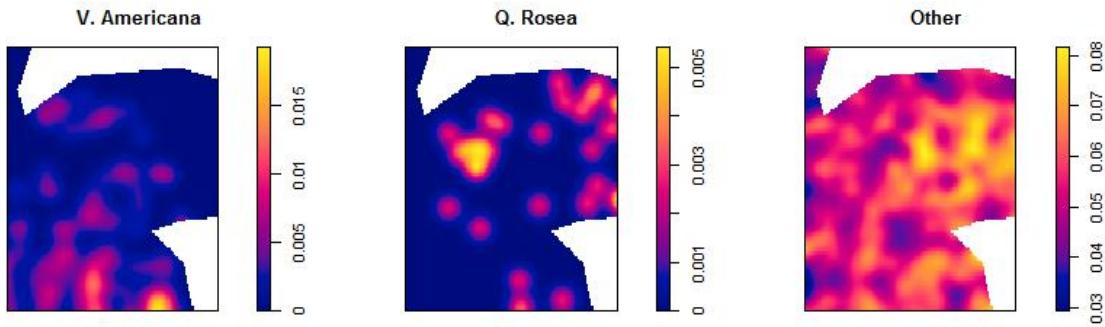


Figure 4.15 – Representation of the density of tree species *Vacapoua americana*, *Qualea rosea* or other (mix) in the Paracou16 forest system.

**Source:** *Paracou16 dataset in package dbmss, authors' calculations*

In the field of spatial economics, the study of multi-type processes could also be rich in information. We could, for example, question the possible interactions between the different types of facilities (general practitioners, schools, etc.). Using the extract from the permanent database of facilities in the city of Rennes, the four spatial sub-distributions were shown in Figure 4.1. In Figure 4.16, we mapped the densities of two facilities: pharmacies and doctors. Visually, quite similar implantation trends seem to be present, as confirmed by the 3D representation in Figure 4.16. The `persp` function in `spatstat` is used.

---

```
library("dbmss")
# BPE file on the INSEE.fr site: https://www.insee.fr
# Data for these examples:
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))

bpe_pharm<- bpe[bpe $TYPEQU=="D301", ]
bpe_docs<- bpe[bpe $TYPEQU=="D201", ]

pharma <- as.ppp(bpe_pharm[,c ("lambert_x", "lambert_y")], owin(c(min(bpe_pharm[, "lambert_x"]),max (bpe_pharm[, "lambert_x"])),c (min(bpe_pharm[, "lambert_y"]),max (bpe_pharm[, "lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
doctors <- as.ppp(bpe_docs[,c ("lambert_x", "lambert_y")], owin(c(min(bpe_docs[, "lambert_x"]),max (bpe_docs[, "lambert_x"])),c (min(bpe_docs[, "lambert_y"]),max (bpe_docs[, "lambert_y"]))))
bpe_doctors_wmppp <- as.wmppp(doctors)

persp(density(doctors),col ="limegreen",
theta = -45,#Viewing angle
xlab = "Lambert X", ylab = "Lambert Y", zlab = "Density",
main = "Doctors")
persp(density(pharma),col ="limegreen", theta = -45,
xlab = "Lambert X", ylab = "Lambert Y", zlab = "Density",
main = "Pharmacies")
```

---

However, only the results of a second order process property analysis will allow us to reach a

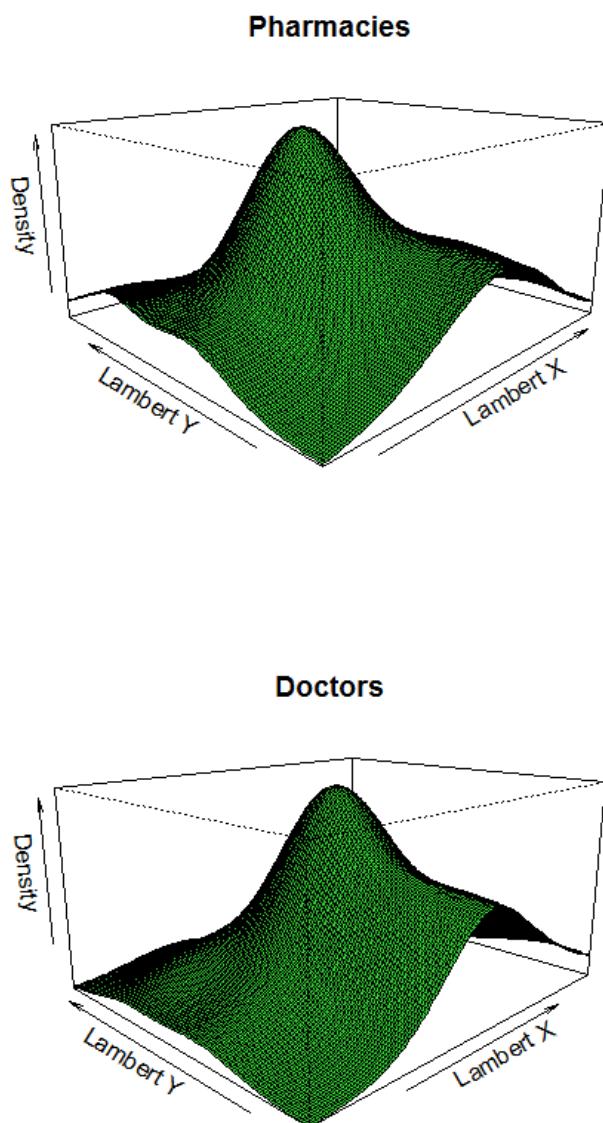


Figure 4.16 – Representation of the density of pharmacies and doctors in Rennes  
**Source:** INSEE-BPE, packages *spatstat* and *dbmss*, authors' calculations

conclusion of any possible interaction (attraction or repulsion) between tree species or between facilities. This is why a first order property study is only a first step of analysis when studying spatial distribution.

#### 4.6.2 Intertype functions

Various developments have been proposed to study the second order properties of multi-type processes. Indicators derived from Ripley's  $K$  function (univariate) have been proposed to analyse relative locations of spatial sub-distributions related to different marks. These indicators are generally referred to as intertype or bivariate functions. We will look at two in more detail in the following sub-sections. From a practical point of view, it is possible to use R packages such as *spatstat* or *dbmss* to calculate the functions and represent the results graphically.

##### The $K$ intertype function

Consider the following case. We would like to study the spatial structure between two types of points, for example:  $T$  type points located around  $S$  type points. Using an intertype function then makes it possible to study the spatial structure of  $T$  type points located at a distance of less than or equal to  $r$  from  $S$  type points.

An initial indicator can be used, the  $K$  intertype function. This is written  $\widehat{K}_{S,T}$  and is defined as follows:

$$\widehat{K}_{S,T}(r) = \frac{1}{\widehat{\lambda}_S n_S} \sum_{i \in S} \sum_{j \in T} \mathbf{1}\{\|x_i - x_j\| \leq r\}. \quad (4.16)$$

where  $\widehat{\lambda}_S$  refers to the estimated intensity of the  $S$  type sub-process. In the field of study,  $n_S$  represents the total number of points  $S$ .

In the event that  $S$  and  $T$  are the same type, the definition of the univariate  $K$  function is presented in the section 4.4.1 (p.83). Note, however, that the correction of edge effects is not included here in the definition of the intertype  $K$  function for ease of presentation. The reference value is always  $\pi r^2$ , regardless of the radius  $r$ , since this is based on the null hypothesis of a completely random distribution of points (of types  $S$  and  $T$ ). If the  $S$  type sub-process is independent of the  $T$  type sub-process, then the number of  $T$  type points within or equal to a distance of  $r$  from an  $S$  type point is the expected number of  $T$  type points located in a disc of radius  $r$ , or  $\lambda_T \pi r^2$ . This null hypothesis corresponds to the independent distribution of two types of industrial establishments, for example. Another null hypothesis giving the same result is that the points are first distributed according to a homogeneous Poisson process and then receive their type in a second stage (for example, commercial spaces are created and then occupied by different types of shops). For all  $r$  distances for which observed values of  $\widehat{K}_{S,T}(r)$  are less than  $\pi r^2$ , a tendency to repulsion of  $T$  points around  $S$  points would be reported. Conversely, values of  $\widehat{K}_{S,T}$  greater than  $\pi r^2$  will tend to validate an attraction of  $T$  points around  $S$  points within a radius  $r$ . The simulation of a confidence interval using the Monte Carlo method will result in an attraction or a repulsion between the two types of points.

The  $K$  intertype function can be implemented in package *spatstat* using the *Kcross* function. In application, let's look again at the example of the Paracou16 data. Indeed, if we use the intertype  $K$  function, we hypothesise that the space in question is homogeneous; however, this hypothesis is almost systematically used in empirical analyses in forest ecology (Goreaud 2000). In Figure 4.17, we have represented the intertype  $K$  functions (or bivariates) for the species *Qualea rosea* or mixed *Other* with that of *Vacapoua americana*. The black curves represent observed  $K$  intertype functions and red dotted lines represent reference intertype  $K$  functions. As can be seen, there is a repulsive relationship between *Qualea rosea* and *Vacapoua americana* (observed  $K$  intertypes are located below the reference value) whereas no association trend appears to exist between the *Vacapoua*

*americana* and other tree species (theoretical and observed  $K$  intertype curves are mixed up for all distances).

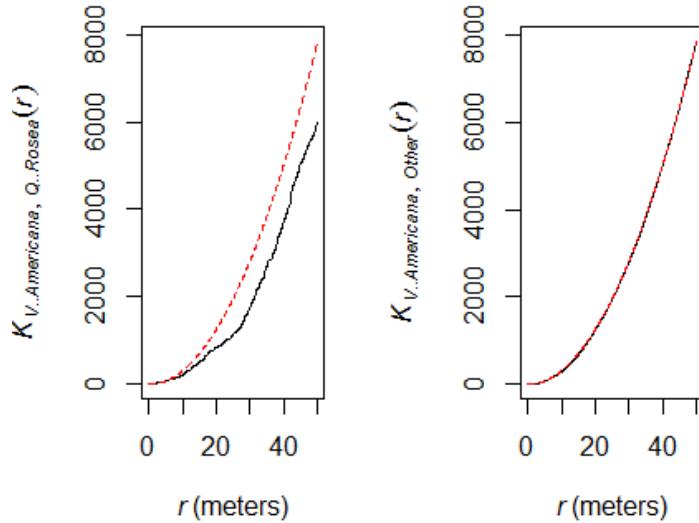


Figure 4.17 – Interactions of different tree species in the Paracou16 forest system

Source: Paracou16 dataset in package dbmss, authors' calculations

---

```
library("dbmss")
# Simplification of marks
marks(paracou16)<- paracou16$marks$PointType
par(mfrow=c(1,2))
# Calculation of K intertypes for the trees of species "Q.Rosea" around
# those of species "Q. Rosea"
plot(Kcross(paracou16, "V. Americana", "Q. Rosea", correction="isotropic"),
     legend=FALSE, main=NULL)
# calculation of K intertypes for trees of species "Q.Rosea" around those
# of species "Other"
plot(Kcross(paracou16, "V. Americana", "Other", correction="isotropic"),
     legend=FALSE, main=NULL)
par(mfrow=c(1,1))
```

---

### The $M$ intertype function

Similarly, the previously presented  $M$  function can be used as an intertype tool. The idea is always to compare a local proportion to a global proportion but in the case of the  $M$  intertype function, the type of neighbouring points of interest is not the same as that of the centre points. For example, if we suspect an attraction of  $T$  type points by  $S$  type points, we will compare the local proportion of  $T$  type neighbours around  $S$  type points to the overall proportion observed throughout the territory in question. If the attraction between the  $T$  type points around  $S$  type points is real, the proportion of  $T$  type points around  $S$  type points should be locally higher than that observed across the entire study area. Conversely, if  $T$  points are repulsed by  $S$  type points, the relative proportion of  $T$  type points around  $S$  type points will be relatively lower than that observed for the whole

territory analysed. In this case, the unweighted empirical estimator of  $M$  intertypes will be defined by:

$$\hat{M}_{S,T}(r) = \sum_{i \in S} \frac{\sum_{j \in T} \mathbf{1}(\|x_i - x_j\| \leq r)}{\sum_{\substack{j \neq i}} \mathbf{1}(\|x_i - x_j\| \leq r)} / \frac{n_T}{n-1}. \quad (4.17)$$

where  $n$  means the total number of points across the entire study area,  $n_S$  the  $S$  type points. As for the intertype  $K$  function, we will assume here that each point belongs to only one type that can be  $S$ ,  $T$  or other. For the intertype  $M$  function, the reference value for all distances  $r$  in question is always equal to 1. For more details on this function (taking into account the weighting, construction of the associated confidence interval, etc.), please refer to the article by Marcon et al. 2010. This intertype function can be calculated in R using the `Mhat` function of package `dbmss`. The `MEvelope` function from the same package can be used to construct a confidence interval.

A concrete example of how to apply  $M$  intertypes is given below, based on the Rennes facilities that were considered in the introduction. If we suspect relationships of attraction or repulsion between several facilities, it is then possible to analyse existing interactions using the intertype  $M$  function. Remember that using the  $M$  function makes it possible to reject the hypothesis of a homogeneous space that can be considered to be a strong hypothesis to characterise the location of economic activities (see, for example, Duranton et al. 2005, p. 1104). In this case, the use of  $M$  intertypes would therefore seem more appropriate than  $K$  intertypes. In Figure 4.18, based on the data extract from the facilities database, we have represented the links between the locations of doctors and pharmacies in Rennes. On the right-hand graphic, the locations of pharmacies in a neighbourhood of  $r$  metres of doctors have been analysed. A repulsion would be detected at very short distances and then intertype aggregation would be observable up to 1 km. The left-hand graphic shows that doctors seem to be relatively agglomerated within a radius of 1 km around the locations of pharmacies in Rennes (the construction of a confidence interval with 100 simulations, for example, would allow us to conclude that the tendency towards dispersion at very short distances is not significant).

---

```
library("dbmss")

# BPE file on the INSEE.fr site: https://www.insee.fr
# Data for these examples:
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))

# Set of marked points
bpe_equip <- bpe[bpe$TYPEQU %in% c("C104", "D301", "B302", "D201"), c(2,3,1)]
colnames(bpe_equip) <- c("X", "Y", "PointType")
bpe_equip_wmppp <- wmppp(bpe_equip)
bpe_pha <- bpe[bpe$TYPEQU=="D301", ]
bpe_doc <- bpe[bpe$TYPEQU=="D201", ]
pharma <- as.ppp(bpe_pha[, c("lambert_x", "lambert_y")], owin(c(min(bpe_pha[, "lambert_x"]),
max(bpe_pha[, "lambert_x"])), c(min(bpe_pha[, "lambert_y"]),
max(bpe_pha[, "lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
doctors <- as.ppp(bpe_doc[, c("lambert_x", "lambert_y")], owin(c(min(bpe_doc[, "lambert_x"]),
max(bpe_doc[, "lambert_x"])), c(min(bpe_med[, "lambert_y"]),
max(bpe_med[, "lambert_y"]))))
bpe_doctors_wmppp <- as.wmppp(doctors)
```

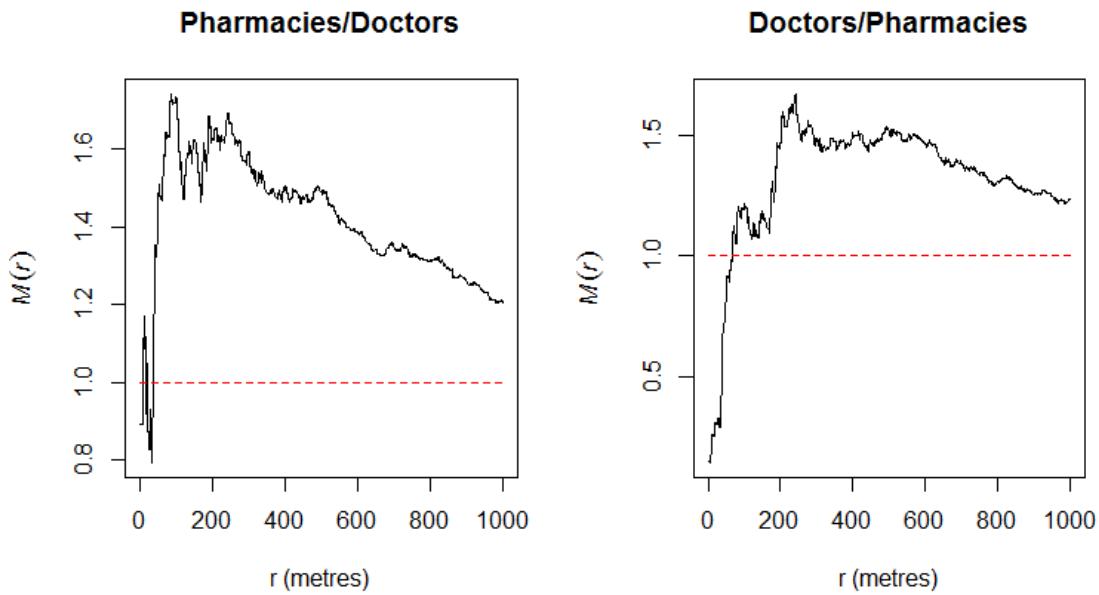


Figure 4.18 – Neighbouring relationships between doctors and pharmacies in Rennes

**Source:** INSEE-BPE, packages *spatstat* and *dbmss*, authors' calculations

```
# Set of marked points
r<- 0:1000

# M intertype: study of interactions between doctors' locations around
# pharmacies
M_pha_doc<- Mhat(bpe_equip_wmppp, r, ReferenceType="D301", NeighborType="D201")

# M intertype: study of interactions between pharmacy locations around
# doctors
M_doc_pha<- Mhat(bpe_equip_wmppp, r, ReferenceType="D201", NeighborType="D301")

par(mfrow=c(1, 2))
plot(M_pha_doc, legend=FALSE, main="Pharmacies/Doctors", xlab = "r (metres")
)
plot(M_doc_pha, legend=FALSE, main="Doctors/Pharmacies", xlab = "r (metres")
)
par(mfrow=c(1, 1))
```

Analysis of the neighbouring relations between Rennes facilities is not the only factor that can be explored. For example, we could suspect interactions between the locations of certain facilities and the population. To examine this relationship, the data in Figure 4.13 would have to be considered with the population data. The R code to establish the link between the population and the four types of facilities considered using the  $M$  function is given below. Figure 4.19 clearly shows that the distribution of the four facilities in question does not appear to deviate significantly

from that of the population (the maximum distance reported was limited to 100 metres as no notable result is obtained beyond this radius).

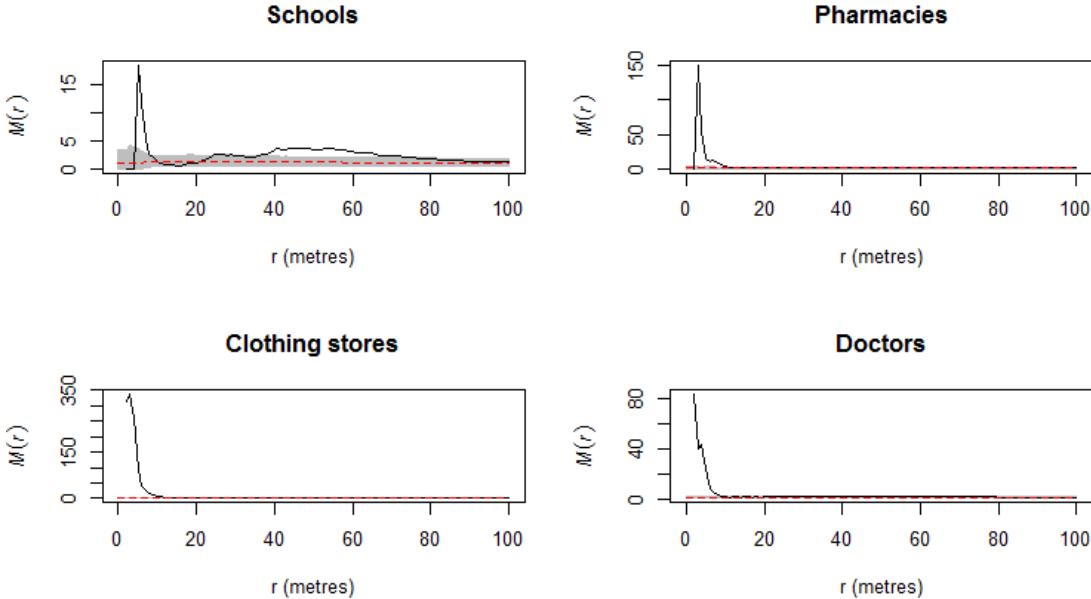


Figure 4.19 – Facilities/population interactions for the four facilities in Rennes

**Source:** INSEE-BPE, packages *spatstat* and *dbmss*, authors' calculations

```

library("dbmss")
colnames(popu) <- c("X", "Y", "PointWeight")
popu$PointType<- "POPU"
popuwmppp<- wmppp(popu)

# Merger of point sets in the window bpe_equip_dbmms
bpe_equip_popu<- superimpose(popuwmppp, bpe_equip_wmppp, W=bpe_equip_wmppp
    $window)

# 100 simulations are selected by default
menv_popu_sch<- MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="C104", SimulationType="RandomLabeling")
menv_popu_ph<- MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="D301", SimulationType="RandomLabeling")
menv_popu_clo<- MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="B302", SimulationType="RandomLabeling")
menv_popu_doc<- MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="D201", SimulationType="RandomLabeling")

par(mfrow=c(2, 2))
plot(menv_popu_sch,legend =FALSE, main="Schools", xlim=c(0,100), xlab = "r
    (metres)")
plot(menv_popu_ph,legend =FALSE, main="Pharmacies", xlim=c(0,100), xlab =
    "r (metres)")

```

---

```
plot(menv_popu_clo, legend=FALSE, main="Clothing stores", xlim=c(0,100),
     xlab = "r (metres)")
plot(menv_popu_doc, legend=FALSE, main="Doctors", xlim=c(0,100), xlab = "r
     (metres)")
par(mfrow=c(1, 1))
```

---

Lastly, note that the  $M$  intertype function is not the only function available in heterogeneous space. Other univariate functions have a bivariate version such as  $K_d$  or  $K_{inhom}$  and can be implemented using package *package dbmss* in R.

## 4.7 Process modelling

The processes presented above, particularly the Poisson processes, are also used to build models. As in traditional statistical models, they are used to explain and predict. The aim is also to find the one with the best power of explanation among the competing models. To build these models, we use covariates. The flexibility of the R software allows the use of data that are associated with observation points, but also continuous data, images and grids.

### 4.7.1 General modelling framework

To adjust a Poisson point process to a spread of points, the shape of the intensity function can be specified  $\lambda(\cdot)$  in order to look for the parameters that allow for the best adjustment. In the *spatstat*, package, the *ppm* function is an essential tool. If we call *trend* the intensity model and *mypp* the process analysed, the command is written:

---

```
ppm(mypp~trend)
# where "trend" refers generically to a trend and
#       "mypp" refers to the process analysed
```

---

The syntax of this point process modelling (PPM) command is similar to that of the standard command *lm* in R, which is used for linear regression models. There are many specifics in modelling: estimated models may result from a log-linear function of the explanatory variable, defined from several variables, etc. The choice and validation of the models must complete the analysis to provide a conclusive response. Among the solutions, the likelihood ratio test may be applied.

### 4.7.2 Application examples

To address such a question, the datasets analysed must be rich enough to satisfy theoretical models. Readers interested in this approach may refer to the two notable examples dealt with in detail in the work of Baddeley et al. 2005. The first is based on data (*Bei*) relating to trees of the species *Beischmiedia pendula* available in package *spatstat*. Indeed, in addition to the location of trees of this species in a tropical rainforest on the island of Barro Colorado, data on the altitude and slope of the land are also provided. The second dataset, named *Murchison* in package *spatstat*, relates to the location of gold deposits in Murchison in West Australia. This is used to model the intensity of gold deposits according to other spatial data: the distance to the nearest geological fault (the faults are described by lines) and the presence of a particular type of rock (described by polygons). Process intensity modelling can therefore be based on exogenous variables that are measured or calculated from geographical information.

Modelling progress is implemented regularly in the *ppm* function. The ability to model interactions between points (with the *interaction* argument in the function) in addition to density currently exists for only one particular type of processes, those of Gibbs, used for the modelling

of the spatial aggregation in industry by Sweeney et al. 2015. The `ppm` function can be used for updates.

## **Conclusion**

In this chapter, we have attempted to give an initial overview of the statistical methods that can be used to characterise point data. Our objective was to emphasise that the diversity of questions raised requires careful handling of statistical tools. Before any study, the question asked and its framework of analysis should, therefore, be clearly defined in order to select the most relevant statistical method. This theoretical warning is important because calculation routines are now widely accessible in the R software in particular and, in principle, pose few practical problems in use. These statistical methods may give rise to more advanced analyses in this field or additional studies, in particular in spatial econometrics for example (see Chapter 6: "Spatial econometrics: current models").

## References - Chapter 4

- Arbia, Giuseppe (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.
- Arbia, Giuseppe, Giuseppe Espa, and Danny Quah (2008). « A class of spatial econometric methods in the empirical analysis of clusters of firms in the space ». *Empirical Economics* 34.1, pp. 81–103.
- Arbia, Giuseppe et al. (2012). « Clusters of firms in an inhomogeneous space: The high-tech industries in Milan ». *Economic Modelling* 29.1, pp. 3–11.
- Baddeley, Adrian J., Jesper Møller, and Rasmus Plenge Waagepetersen (2000). « Non- and semi-parametric estimation of interaction in inhomogeneous point patterns ». *Statistica Neerlandica* 54.3, pp. 329–350.
- Baddeley, Adrian J., Edge Rubak, and R Turner (2015b). *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics. 810 pages. Chapman and Hall/CRC.
- Baddeley, Adrian J and Rolf Turner (2005). « Spatstat: an R package for analyzing spatial point patterns ». *Journal of Statistical Software* 12.6, pp. 1–42.
- Barlet, Muriel, Anthony Briant, and Laure Crusson (2008). *Concentration géographique dans l'industrie manufacturière et dans les services en France : une approche par un indicateur en continu*. Série des documents de travail de la Direction des Études et Synthèses économiques G 2008 / 09. Institut National de la Statistique et des études économiques (Insee).
- (2013). « Location patterns of service industries in France: A distance-based approach ». *Regional Science and Urban Economics* 43.2, pp. 338–351.
- Behrens, Kristian and Théophile Bougna (2015). « An anatomy of the geographical concentration of Canadian manufacturing industries ». *Regional Science and Urban Economics* 51, pp. 47–69.
- Besag, Julian E. (1977). « Comments on Ripley's paper ». *Journal of the Royal Statistical Society B* 39.2, pp. 193–195.
- Bonneu, Florent (2007). « Exploring and Modeling Fire Department Emergencies with a Spatio-Temporal Marked Point Process ». *Case Studies in Business, Industry and Government Statistics* 1.2, pp. 139–152.
- Bonneu, Florent and Christine Thomas-Agnan (2015). « Measuring and Testing Spatial Mass Concentration with Micro-geographic Data ». *Spatial Economic Analysis* 10.3, pp. 289–316.
- Briant, Anthony, Pierre-Philippe Combes, and Miren Lafourcade (2010). « Dots to boxes: Do the Size and Shape of Spatial Units Jeopardize Economic Geography Estimations? » *Journal of Urban Economics* 67.3, pp. 287–302.
- Brülhart, Marius and Rolf Traeger (2005). « An Account of Geographic Concentration Patterns in Europe ». *Regional Science and Urban Economics* 35.6, pp. 597–624.
- Cole, Russel G. and Gregg Syms (1999). « Using spatial pattern analysis to distinguish causes of mortality: an example from kelp in north-eastern New Zealand ». *Journal of Ecology* 87.6, pp. 963–972.
- Combes, Pierre-Philippe, Thierry Mayer, and Jacques-François Thisse (2008). *Economic Geography, The Integration of Regions and Nations*. Princeton: Princeton University Press.
- Combes, Pierre-Philippe and Henry G Overman (2004). « The spatial distribution of economic activities in the European Union ». *Handbook of Urban and Regional Economics*. Ed. by J Vernon Henderson and Jacques-François Thisse. Vol. 4. Amsterdam: Elsevier. North Holland. Chap. 64, pp. 2845–2909.
- Condit, Richard et al. (2000). « Spatial Patterns in the Distribution of Tropical Tree Species ». *Science* 288.5470, pp. 1414–1418.
- Diggle, Peter J. (1983). *Statistical analysis of spatial point patterns*. London: Academic Press, 148 p.

- Diggle, Peter J. and A. G. Chetwynd (1991). « Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations ». *Biometrics* 47.3, pp. 1155–1163.
- Duranton, Gilles (2008). « Spatial Economics ». *The New Palgrave Dictionary of Economics*. Ed. by Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan.
- Duranton, Gilles and Henry G. Overman (2005). « Testing for Localization Using Micro-Geographic Data ». *Review of Economic Studies* 72.4, pp. 1077–1106.
- Ellison, Glenn and Edward L. Glaeser (1997). « Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach ». *Journal of Political Economy* 105.5, pp. 889–927.
- Ellison, Glenn, Edward L. Glaeser, and William R. Kerr (2010). « What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns ». *The American Economic Review* 100.3, pp. 1195–1213.
- Fehmi, Jeffrey S. and James W. Bartolome (2001). « A grid-based method for sampling and analysing spatially ambiguous plants. » *Journal of Vegetation Science* 12.4, pp. 467–472.
- Goreaud, François and Raphaël Pélissier (1999). « On explicit formulas of edge effect correction for Ripley's K-function ». *Journal of Vegetation Science* 10.3, pp. 433–438. ISSN: 1654-1103. DOI: 10.2307/3237072. URL: <http://dx.doi.org/10.2307/3237072>.
- Goreaud, François (2000). « Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes ». PhD Thesis. Nancy: ENGREF.
- Heinrich, Lothar (1991). « Goodness-of-fit tests for the second moment function of a stationary multidimensional poisson process ». *Statistics: A Journal of Theoretical and Applied Statistics* 22.2, pp. 245–268. DOI: 10.1080/02331889108802308.
- Holmes, Thomas J. and John J. Stevens (2004). « Spatial Distribution of Economic Activities in North America ». *Cities and Geography*. Ed. by J. Vernon Henderson and Jacques-François Thisse. Vol. 4. Handbook of Regional and Urban Economics Chapter 63 - Supplement C. Elsevier, pp. 2797–2843.
- Illian, Janine et al. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Statistics in Practice. Chichester: Wiley-Interscience, p. 534.
- Jensen, Pablo and Julien Michel (2011). « Measuring spatial dispersion: exact results on the variance of random spatial distributions ». *The Annals of Regional Science* 47.1, pp. 81–110.
- Lagache, Thibault et al. (2013). « Analysis of the Spatial Organization of Molecules with Robust Statistics ». *Plos One* 8.12, e80914.
- Lang, G., E. Marcon, and F. Puech (2015). « Distance-Based Measures of Spatial Concentration: Introducing a Relative Density Function ». *HAL hal-01082178.version 2*.
- Lang, Gabriel and Eric Marcon (2013). « Testing randomness of spatial point patterns with the Ripley statistic ». *ESAIM: Probability and Statistics* 17, pp. 767–788.
- Marcon, Eric and Florence Puech (2003). « Evaluating the Geographic Concentration of Industries Using Distance-Based Methods ». *Journal of Economic Geography* 3.4, pp. 409–428.
- (2010). « Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods ». *Journal of Economic Geography* 10.5, pp. 745–762.
- (2015a). « Mesures de la concentration spatiale en espace continu : théorie et applications ». *Économie et Statistique* 474, pp. 105–131.
- (2017). « A typology of distance-based measures of spatial concentration ». *Regional Science and Urban Economics* 62, pp. 56–67.
- Marcon, Eric, Florence Puech, and Stéphane Traissac (2012). « Characterizing the relative spatial structure of point patterns ». *International Journal of Ecology* 2012. Article ID 619281, p. 11.
- Marcon, Eric et al. (2015b). « Tools to Characterize Point Patterns: dbmss for R ». *Journal of Statistical Software* 67.3, pp. 1–15.
- Maurel, Françoise and Béatrice Sédillot (1999). « A measure of the geographic concentration in french manufacturing industries ». *Regional Science and Urban Economics* 29.5, pp. 575–604.

- Møller, Jesper and Hakon Toftaker (2014). « Geometric Anisotropic Spatial Point Pattern Analysis and Cox Processes ». *Scandinavian Journal of Statistics*. Monographs on Statistics and Applies Probabilities 41.2, pp. 414–435.
- Møller, Jesper and Rasmus Plenge Waagepetersen (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Vol. 100. Monographs on Statistics and Applies Probabilities. Chapman and Hall, 300 p.
- Openshaw, S. and P. J. Taylor (1979a). « A million or so correlation coefficients: three experiments on the modifiable areal unit problem ». *Statistical Applications in the Spatial Sciences*. Ed. by N. Wrigley. London: Pion, pp. 127–144.
- Ripley, Brian D. (1976). « The Second-Order Analysis of Stationary Point Processes ». *Journal of Applied Probability* 13.2, pp. 255–266.
- (1977). « Modelling Spatial Patterns ». *Journal of the Royal Statistical Society B* 39.2, pp. 172–212.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall, 175 p.
- Sweeney, Stuart H. and Edward J. Feser (1998). « Plant Size and Clustering of Manufacturing Activity ». *Geographical Analysis* 30.1, pp. 45–64.
- Sweeney, Stuart H and Miguel Gómez-Antonio (2015). « Localization and Industry Clustering Econometrics: an Assessment of Gibbs Models for Spatial Point Processes ». *Journal of Regional Science* 56.2, pp. 257–287.
- Szwagrzyk, Jerzy and Marek Czerwczak (1993). « Spatial patterns of trees in natural forests of East-Central Europe ». *Journal of Vegetation Science* 4.4, pp. 469–476.
- Veen, Alejandro and Frederic Paik Schoenberg (2006). « Assessing Spatial Point Process Models Using Weighted K-functions: Analysis of California Earthquakes ». *Case Studies in Spatial Point Process Modeling*. Ed. by Adrian Baddeley et al. New York, NY: Springer New York, pp. 293–306.
- Wiegand, T. and K. A. Moloney (2004). « Rings, circles, and null-models for point pattern analysis in ecology ». *Oikos* 104.2, pp. 209–229.



# 5. Geostatistics

JEAN-MICHEL FLOCH

INSEE

---

<b>5.1</b>	<b>Random functions</b>	<b>114</b>
5.1.1	Definitions . . . . .	114
5.1.2	Stationarity . . . . .	115
<b>5.2</b>	<b>Spatial variability</b>	<b>116</b>
5.2.1	Covariance and correlogram . . . . .	116
5.2.2	Variogram . . . . .	117
5.2.3	Empirical application . . . . .	117
<b>5.3</b>	<b>Fitting variogram</b>	<b>122</b>
5.3.1	General shape of the variogram . . . . .	122
5.3.2	Usual variograms . . . . .	123
5.3.3	Variogram fitting . . . . .	125
<b>5.4</b>	<b>Ordinary kriging</b>	<b>127</b>
5.4.1	Principle . . . . .	127
5.4.2	Application to rainfall data . . . . .	129
<b>5.5</b>	<b>Support and change of support</b>	<b>134</b>
5.5.1	Empirical dispersion variance and Krige additivity relationships . . . . .	134
5.5.2	Variogram of the regularised variable . . . . .	134
5.5.3	Block kriging . . . . .	136
<b>5.6</b>	<b>Extensions</b>	<b>136</b>
5.6.1	Cokriging . . . . .	136
5.6.2	Universal kriging . . . . .	139
<b>5.7</b>	<b>Combined models with variogram</b>	<b>141</b>

---

## Abstract

Geostatistics is a very important branch of spatial statistics. It has been developed on the basis of very practical concerns (mining research), and has undergone very significant methodological developments driven by Georges Matheron and his colleagues at the Fontainebleau Ecole des Mines. The simplest illustrations relate to problems such as the interpolation of temperatures or precipitation. But the most important work concerns geological and mining applications (*e.g.* Chiles et al. 2009). Applying geostatistics to demographic or social examples is more difficult, but it seems important to present the broad outline of the method — how to treat stationarity and introduce intrinsic stationarity; introducing the semivariogram to study spatial relations; data interpolation using the kriging method. Beyond mining applications, variogram analysis can be used in mixed models to analyse residuals.

**R** Prior reading of Chapter 1: "Descriptive spatial analysis", and Chapter 4: "Spatial distributions of points" is recommended.

Modelling spatial data is made difficult by the fact that there is only one realisation of the phenomenon. As in spatial point patterns, only one realisation is also observed, for which all the data are available. For continuous data (potentially observable at any point in the space), only partial data are available, from which values at unobserved points may be predicted. It is this lack of information that will lead to the use of probabilistic models.

Randomness is not a property of the phenomenon, but a characteristic of the model used to describe it. Geostatistics, which studies continuous phenomena, has enabled the development of specific methods to study spatial relationships between observations and to construct predictive tools.

Geostatistics owes its name to the discipline's origins in mining (Krig, Matheron). Many of the basic concepts of the discipline derive from the work of Georges Matheron (*regionalized variable*, *random function*, *intrinsic assumption*, *nugget effect*<sup>1</sup>, Matheron et al. 1965). Figure 5.1 presents a summary illustrating the pathway from reality towards a more abstract model, a model that will itself enable us to act in expected best way (Chauvet 2008).

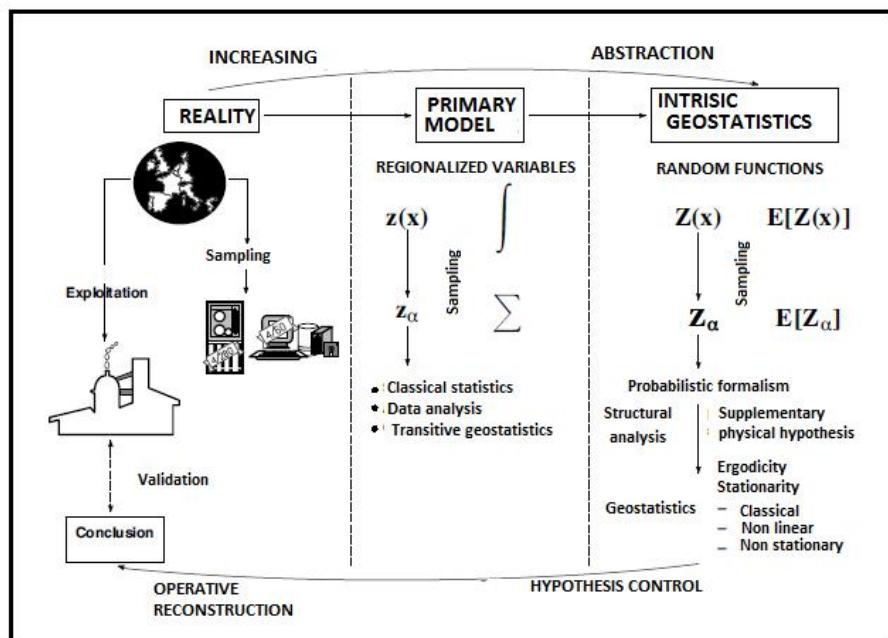


Figure 5.1 – Diagram of a geostatistical analysis

Source: Chauvet 2008

## 5.1 Random functions

### 5.1.1 Definitions

As in Figure 5.1, we designate  $z(s)$  the *regionalized variable*, and  $Z(s)$  the *random function*, the letter  $s$  designating the position in space. We will retain this wording here as specific to geostatistics. A phenomenon occurring in space is qualified as regionalized. A regionalized variable is a function that describes this phenomenon satisfactorily. This is a first level of abstraction, where we remain

1. These terms are defined later in this chapter.

in the description, without resorting to a probabilistic model. If we make no additional assumption, we remain within the framework of *transitive geostatistics*.

The next step, qualified as *intrinsic geostatistics*, introduces the concept of random function. It results from a choice — considering the regionalized variable as the realisation of a random function. This choice makes it possible to use powerful probabilistic tools, the counterpart being moving further from reality. The probabilistic model is a calculation intermediary that is expected to be used in understanding the regionalized phenomenon.

The random function is fully characterised by the knowledge of its distribution function.

$$F(s_1, s_2, \dots, s_n; z_1, z_2, \dots, z_n) = P\{Z(s_1) \leq z_1, Z(s_2) \leq z_2, \dots, Z(s_n) \leq z_n\}. \quad (5.1)$$

Since we have only one realisation of our regionalized phenomenon, we need to find another way to make the inference. Citing Matheron et al. 1965: "For inference to be possible, we need to introduce additional assumptions about random function  $Z(s)$  so as to reduce the number of parameters on which its law depends. This is the purpose of the stationary assumption we are going to define: a stationary function is repeated itself in some form in space, and this repetition once again makes statistical inference possible from a single realisation." Each observation will therefore be treated as the realisation of a random variable.

### 5.1.2 Stationarity

Three meanings of stationarity are used in geostatistics:

- strict stationarity;
- second order stationarity;
- intrinsic stationarity.

**Definition 5.1.1 — Strict stationarity.** Strict stationarity directly refers to the probability law of the process. There is strict stationarity if by moving using translation, all the characteristics of the random function remain the same.

Formally, the joint distribution of  $Z(s_i)$  is the same as that of  $Z(s_i + h)$ ,  $h$  indicating translation relative to the initial position. This form of stationarity is not operational and very restrictive.

**Definition 5.1.2 — Second order stationarity.** Second order stationarity or weak stationarity no longer imposes conditions on the probability law, but only on the mean and covariance. These indicators must be invariable by translation.

Given that  $Z(s)$  breaks down into a deterministic component and a random component  
 $Z(s) = m(s) + R(s)$

second order stationarity requires the following conditions:

- $E[Z(s)] = m(s) \forall s$ .  
The invariance of the expected value by translation leads to constancy of the deterministic component.  
 $m(s+h) = m(s) = m \forall s$ ;
- The variance is constant:  $E[(Z(s) - m)^2] = \sigma^2$ ;
- Covariance depends only on the spatial shift:  
 $Cov[Z(s+h), Z(s)] = E[(Z(s+h) - m)(Z(s) - m)] = C(h)$ .

In practice, this stationarity assumption is often too strong. The most important limit is that the mean can change over the area of interest, and that the variance may not be bounded when this area of interest grows. It was George Matheron who drew the consequences of the limits of weak stationarity by suggesting the even weaker concept of intrinsic stationarity (Matheron et al. 1965).

**Definition 5.1.3 — Intrinsic stationarity.** The intrinsic stationarity assumption is as follows:

$$E[(Z(s+h) - Z(s))^2] = 0.$$

Increments can be stationary without the process itself being stationary.

A new function can then be defined, called *variogram*, based on differences between values and shifted values, and which depends only on the offset:

$$\gamma(h) = \frac{1}{2} E[Z(s+h) - Z(s)]^2 \quad (5.2)$$

Second order stationarity leads to intrinsic stationarity, but the reverse is not true. A random function can enable a variogram to be calculated, but this is not the case for covariance and the autocorrelation function.

## 5.2 Spatial variability

### 5.2.1 Covariance and correlogram

**Definition 5.2.1 — Covariance.** The covariance function will allow the relationships between all point pairs to be considered. If we consider two points  $s_i$  and  $s_j$ , covariance can be defined by Equation 5.3.

$$\text{Cov}[Z(s_i), Z(s_j)] = E[(Z(s_i) - m)(Z(s_j) - m)] \quad (5.3)$$

When the process is second-order stationary, covariance will no longer depend only on the distance between the points  $|s_i - s_j|$ . If we designate  $h$  this distance, we will define  $C(h)$  calculated for all values of  $h$  taking into account all point pairs located at a distance  $h$  from each other. Covariance function  $C(h)$  is defined by Equation 5.4.

$$C(h) = \text{Cov}[Z(s+h), Z(s)] = E[(Z(s+h) - m)(Z(s) - m)] \quad (5.4)$$

It reflects how the covariance of observations changes when their distance increases. When  $h$  is equal to 0, covariance is equal to variance.

$$C(0) = E[(Z(s) - m)^2] = \sigma^2 \quad (5.5)$$

The covariance function has the following properties:

$$C(-h) = C(h) \quad (5.6)$$

$$|C(h)| \leq C(0). \quad (5.7)$$

For the covariance function to be called *allowable*, the variance of a linear combination of variables must be positive:  $\text{Var}[\sum_{i=1}^n \lambda_i z(s_i)] = \sum_{i=1}^n \sum_{j=1}^n C(s_i - s_j)$ .

This results in  $C$  being *positive semi-definite*.

**Definition 5.2.2 — Autocorrelation function.** The autocorrelation function  $\rho(h)$  is defined as a function of  $h$  by ratio  $\frac{C(h)}{C(0)}$ . Its value is between -1 and +1. The following relationships can be shown when second order stationarity is verified:

$$\begin{aligned} \gamma(h) &= C(0) - C(h) \\ \gamma(h) &= \sigma^2 (1 - \rho(h)). \end{aligned} \quad (5.8)$$

**Box 5.2.1 — Estimate of the covariance function.** The covariance function is estimated from  $n(h)$  point pairs, as defined below, for  $i$  variant from 1 to  $n(h)$ .

$$\hat{C}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(s_i) - m)(z(s_i + h) - m) \quad (5.9)$$

with  $n(h) = \text{Card}\{(s_i, s_j) / |s_i - s_j| \approx h\}$

### 5.2.2 Variogram

The literature contains *variogram* or *semivariogram* expressions. Some authors (Matheron et al. 1965) believe that the term semivariogram should be used for  $\gamma(h)$  as defined in Equation 5.10, the variogram corresponding to  $2\gamma(h)$ . This is the choice we make in this article.

Out of the following three indicators – covariance function, autocorrelation function and variogram – the latter is most used to the extent that it refers to the weakest form of stationarity and therefore to the least restrictive conditions on the local behaviour of the mean.

$$\gamma(h) = \sigma^2(1 - \rho(h)) \quad (5.10)$$

The variogram has the following properties:

$$\begin{aligned} \gamma(h) &= \gamma(-h) \\ \gamma(0) &= 0 \\ \frac{\gamma(h)}{\|h\|^2} &\rightarrow 0 \quad \text{quand} \quad \|h\| \rightarrow \infty \end{aligned} \quad (5.11)$$

For any set of real  $\{a_1, a_2, \dots, a_m\}$  verifying  $\sum_{i=1}^m a_i = 0$ , we have the following property:

$$\sum_{i=1}^m \sum_{j=1}^m a_i a_j \gamma(s_i - s_j) \leq 0 \quad (5.12)$$

When the process is isotropic:

$$\gamma(h) = \gamma(\|h\|) \quad (5.13)$$

**Box 5.2.2 — Estimate of the experimental variogram.** An experimental variogram can be estimated from point pairs defined as previously.

$$\hat{\gamma}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(s_i + h) - z(s_i))^2 \quad (5.14)$$

where  $n(h) = \text{Card}\{(s_i, s_j) / |s_i - s_j| \approx h\}$

The variogram can be estimated in different directions to highlight any anisotropy of the phenomenon studied.

### 5.2.3 Empirical application

#### Geostatistics with R

The packages dedicated to geostatistics in this chapter are **gstat** and **geoR**, which are most commonly used. Many other packages are available on the CRAN website. Below is a list with comments produced by Roger Bivand.

Package **gstat** provides a wide range of functions for univariate and multivariate geostatistics, including larger data sets, while **geoR** and **geoRglm** contain functions for model-based geostatistics. Variogram diagnostics can be performed with **vardiag**. Automated interpolation using gstat is available in **automap**. This family of packages is supplemented by **intamap** with automatic interpolation procedures and **psgp**, which implements Gaussian process regression kriging. A broad range of similar functions is found in the **fields** package. The spatial package is delivered with the R database and contains several main functions. The **spBayes** package is compatible with Gaussian univariate and multivariate models with MCMC. The **rampes** package is another Bayesian geostatistical modelling set. The **geospt** package contains basic geostatistical and radial functions, including prediction and cross-validation. In addition, it includes functions to design optimal spatial sampling networks based on geostatistical modelling. **spsann** is another package that offers functions to optimise sample configurations, using spatial simulated annealing. The **geostatsp** package offers geostatistical modelling functions using Raster and SpatialPoints objects. Non-Gaussian models are adapted using INLA, and Gaussian geostatistical models use the estimate of maximum likelihood. The **FRK** package is a spatial/spatial-temporal modelling and prediction tool with large data sets. The approach, discussed in Cressie and Johannesson (2008), breaks down the field, and therefore the covariance function, using a fixed set of n basic functions, where n is generally much smaller than the number of data points (or polygons).

### RGeostats package

RGeostats is an R-language package. It was developed by the geostatistics team at the Geosciences Centre of Mines ParisTech. It implements all geostatistical functions available from the (commercial) Geoslib library (written in C/C++). It therefore benefits from the experience accumulated in the mining field, and is especially dedicated to such applications. It makes it possible to carry out all the implementations described in this chapter, and to deal with the support effects.

RGeostats allows R users, by loading and installing it, to access the normal Geostatistics functionalities. It is also a platform that enables the Geostatistics team at the Geosciences Centre to develop prototypes for the application of new models (*e.g.* Boolean simulations, bi-plurigaussian simulations) or new techniques (flow of fluids, simulations of the first arrival time in geophysics). The functions are described, but the code is not accessible, and there are few examples.

RGeostats can be downloaded from the following address: [cg.ensmp.fr/rgeostats](http://cg.ensmp.fr/rgeostats).

### Exploratory analyses

The proposed application uses *Swiss rainfall* data. This database is widely used in spatial studies, particularly in Diggle et al. 2003. It is supplied in the R *geoR* package. The observations are rainfall readings from 467 weather stations in Switzerland, and were made on 8 May, 1986. This is indeed a continuous data point, as rainfall can potentially be recorded at any point in the country. It can therefore fall within geostatistical modelling, but it is easier to understand than mining data. In addition to rainfall, measured in millimetres, data about the altitude of weather stations is provided.

In the *geoR* package, there are three Spatial Interpolation Comparison (SIC) databases:

- *Sic.100*: sample of 100 observations that may be used to make interpolations;
- *Sic.367*: observations not included in the sample that will enable estimates and observations to be compared;
- *Sic.all*: set.

The features of *geoR* will be used here, but *gstat* and *RGeostats* provide analysis tools.

Figure 5.2 shows the data sampled (in green) and the control data, the circles being proportional to the rainfall recorded.

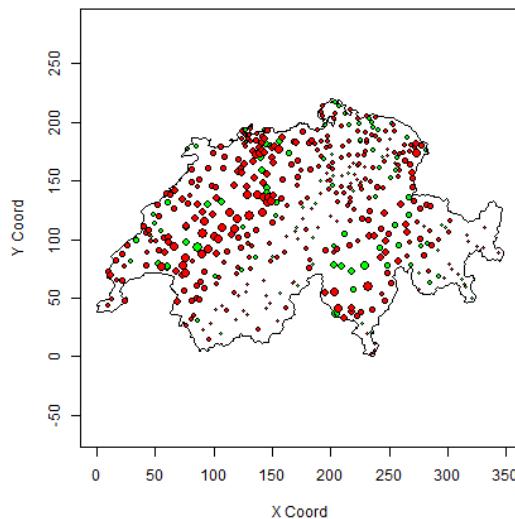


Figure 5.2 – Rainfall in Switzerland

**Source:** Swiss rainfall from the *geoR* package

---

```
library(geoR)
points(sic.100, borders=sic.borders, col="green")
points(sic.367, borders=sic.borders, col="red", add=TRUE)
```

---

The *geoR* package provides some descriptive representations using the `plotgeodata` function. Figure 5.3 shows, from left to right and top to bottom:

- representation of the level of rainfall, based on the quantiles of the variable;
- rainfall based on latitude;
- rainfall based on longitude;
- the histogram of rainfall data.

---

```
library(geoR)
plot.geodata(sic.100, bor=sic.borders)
```

---

The histogram in Figure 5.3 suggests that the variable's distribution is not Gaussian, and that a data transformation could be considered since the most common methods only have interesting properties in the Gaussian context.

### Variogram cloud and experimental variogram

In intrinsic geostatistics, the **variogram cloud** is a cloud of data points expressing their variability based on their interspacing. The variogram cloud provides the graphical representation of the values used to calculate the variogram. For a dataset of variable  $Z$  at points  $(s_1, \dots, s_i, \dots, s_n)$ , it represents the abscissa points  $\|s_i - s_j\|$  and ordinate points  $\frac{1}{2} [z(s_i) - z(s_j)]^2$ . It can be represented as a point cloud and as a boxplot (see Figure 5.4).

---

```
library(geoR)
library(fields)
vario.b<- variog(sic.100, option =c ("bin", "cloud", "smooth"),
bin.cloud=TRUE)
```

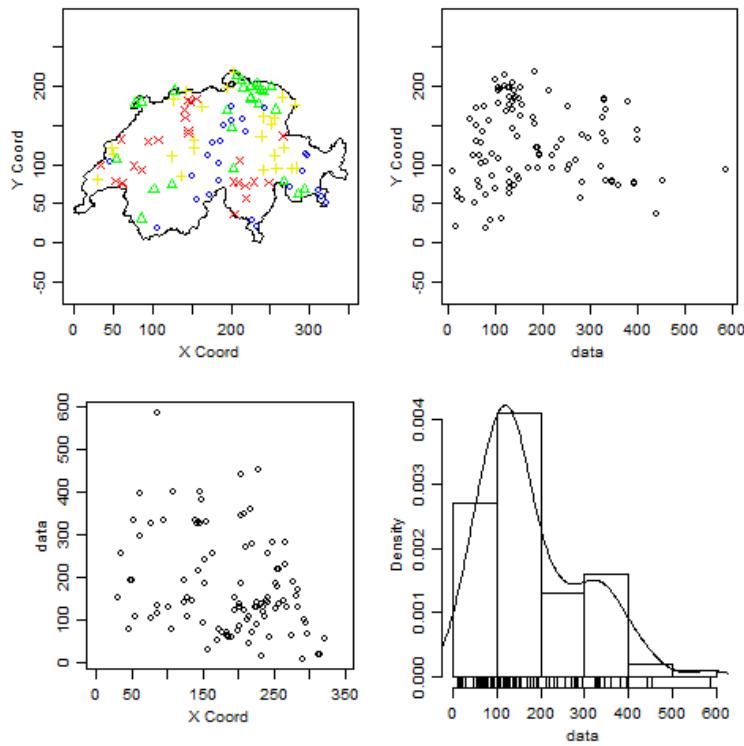


Figure 5.3 – Some descriptive statistics

**Source:** Swiss rainfall from the geoR package

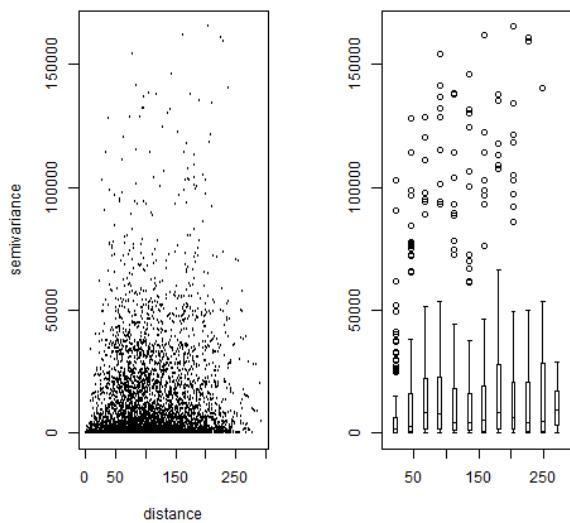


Figure 5.4 – Variogram cloud

**Source:** Swiss rainfall from the geoR package

---

```
vario.c <- variog(sic.100, op="cloud")
bplot.xy(vario.c$u,vario.c$v, breaks=vario.b$u,col="grey80",
lwd=2,cex=0.1,outline=FALSE)
```

---

Since these representations are hard to read, the most useful representation is the experimental variogram (defined in box 5.2.2), shown in Figure 5.6 based on a construction diagram shown in Figure 5.5.

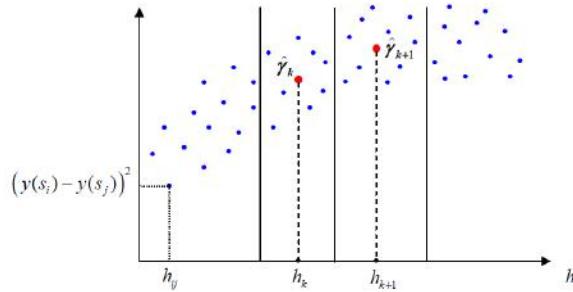


Figure 5.5 – Experimental variogram: construction diagram

**Source:** Swiss rainfall from the *geoR* package

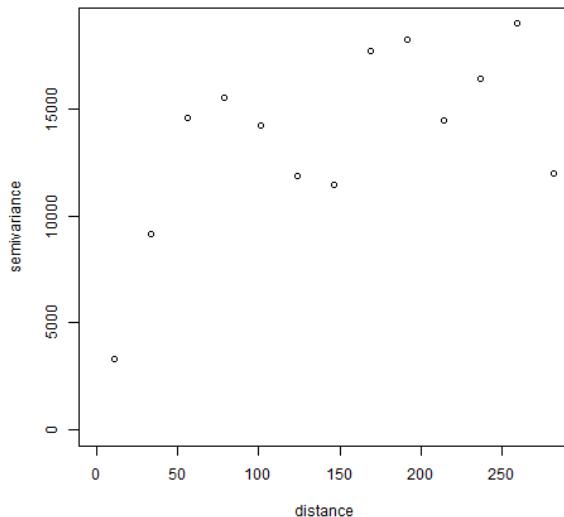


Figure 5.6 – Experimental variogram

**Source:** Swiss rainfall from the *geoR* package

---

```
library(geoR)
vario.ex<- variog(sic.100, bin.cloud=TRUE)
plot(vario.ex)
```

---

In Figure 5.6, all observed points are used to calculate the variogram. But the phenomena studied are not necessarily isotropic, and it may be useful to calculate variograms based on several directions of space (see Figure 5.7).

---

```
library(geoR)
```

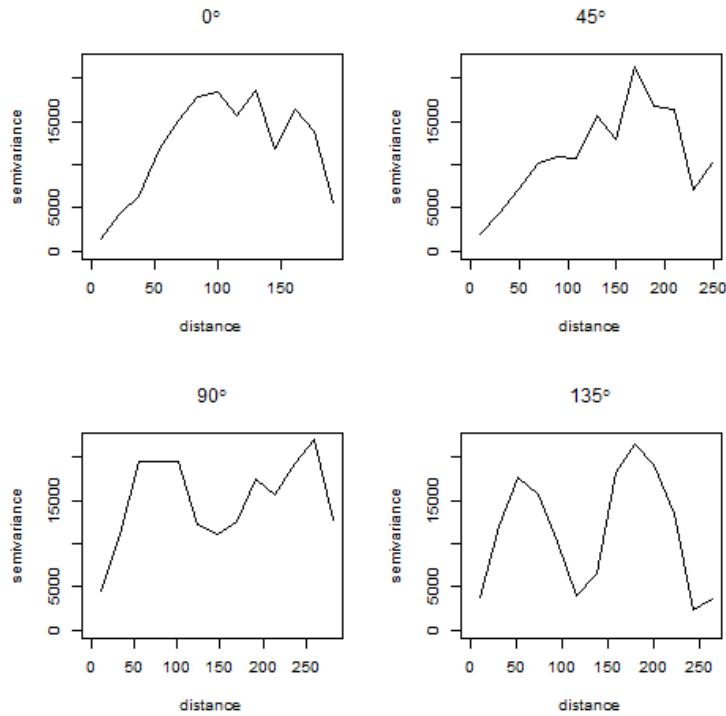


Figure 5.7 – Directional variograms

**Source:** Swiss rainfall from the *geoR* package

```
vario4<-variog4(sic.100)
plot(vario4,same=FALSE)
```

### 5.3 Fitting variogram

In section 5.4, which is dedicated to kriging, we see that the value of the estimators depends on observations and the spatial autocorrelation structure, understood by the variogram. Variogram analysis is therefore not just a passage point. It forms the central point of the geostatistical approach. The empirical variograms shown in section 5.2 cannot be used directly because they don't meet the properties listed in section 5.2.2. To be used in geostatistical models, they must first be adjusted to theoretical models with well-defined analytical forms, which implies a vision of what a semi-variogram should be.

#### 5.3.1 General shape of the variogram

We begin by presenting the most classical model, from which theoretical variograms will be constructed, making it possible to define, among other things, the kriging equations (see section 5.4).

This variogram has a form that is initially increasing, up to a certain level. The value of  $h$  corresponding to this plateau is called the *range*. We understand it by referring to the relationship between covariance, when it is defined, and the semi-variogram, *i.e.*  $\gamma(h) = C(0) - C(h)$ . Covariance is very often a decreasing function of distance, which implies the increase of the semi-variogram, but this is not always the case (*e.g.* cardinal sine model).

At a certain distance, which is called the range, covariance will be cancelled out. This range is the range of spatial dependence. There is no longer any relationship between the values observed at

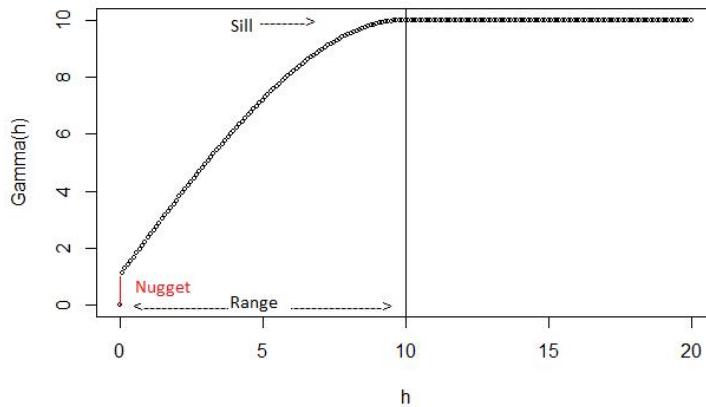


Figure 5.8 – Theoretical variogram

a distance beyond this range. For the semi-variogram this means that, beyond this range, its value is constant and we have:  $C = C(0) = \sigma^2$ .

For  $h = 0$ , the value of the variogram is zero, by definition. But in practice we see that for values very close to 0, the variogram takes values greater than 0 and there is therefore a discontinuity at the origin. We call the limit of the variogram at zero a *nugget*. As Matheron explains (Matheron et al. 1965): "The concept of scale plays a key role here. At a scale of ten metres, a transition phenomenon where the range is measured in centimetres is only seen on  $\gamma(h)$  as a discontinuity at the origin, i.e. a nugget effect." It represents the variation between two measurements made at infinitely close locations, and from two effects there may arise:

- variability of the measuring instrument: the nugget therefore partly measures the statistical error of the measuring instrument.
- a real nugget effect: a sudden change in the measured parameter; the historical case is the passage without transition from a gold nugget to soil containing virtually no gold.

Other forms of variograms may be encountered. Figure 5.9 shows two classical cases. The first is the **linear variogram**, the second the **pure nugget effect**. When the variogram is unbounded, the mean and variance are not defined. This most frequently indicates a large-scale trend, which must be modelled. The pure nugget effect reflects the lack of spatial dependence.

### 5.3.2 Usual variograms

Geostatistical literature offers many functions that satisfy the properties of the semi-variogram as shown in Figure 5.8. These configured functions must be used to describe the different components (range, plateau, nugget). They must also handle the behaviour of the function at the origin (linear trend, horizontal or vertical tangent).

We will only discuss four examples of variogram models here (Figure 5.10), the others being described in the reference works (Armstrong 1998, Chiles et al. 2009, Waller et al. 2004).

#### Definition 5.3.1 — Spherical model.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s \left[ \frac{3}{2} \left( \frac{h}{a} \right) - \frac{1}{2} \left( \frac{h}{a} \right)^3 \right] & 0 < h \leq a \\ c_0 + c_s & h > a \end{cases} \quad (5.15)$$

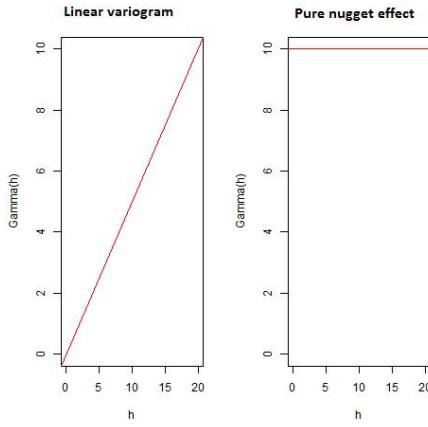


Figure 5.9 – Two atypical semi-variograms

**Definition 5.3.2 — Exponential model.**

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s [1 - \exp(-\frac{h}{a})] & h > 0 \end{cases} \quad (5.16)$$

**Definition 5.3.3 — Gaussian model.**

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s [1 - \exp(-(\frac{h}{a})^2)] & h > 0 \end{cases} \quad (5.17)$$

**Definition 5.3.4 — Power model.**

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + b h^p & h > 0 \end{cases} \quad (5.18)$$

**Definition 5.3.5 — Matern model.**

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_s \left[ 1 - \frac{h}{2^{\alpha-1} \Gamma(\alpha)} K_{\alpha} \left( \frac{h}{a} \right) \right] & h > 0 \end{cases} \quad (5.19)$$

where  $\Gamma$  refers to the gamma function and  $K_{\alpha}$ , the modified Bessel of the second kind of parameter  $\alpha$ .

**Definition 5.3.6 — Cardinal sine model.**

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_s \left[ 1 - \frac{a}{h} \sin \left( \frac{h}{a} \right) \right] & h > 0 \end{cases} \quad (5.20)$$

As explained by Matheron and the geostatisticians of the Ecole des Mines in Fontainebleau, modelling is a matter of choice. Choosing a theoretical model is a decisive moment in the geostatistician's approach, but we cannot associate *a priori* a theoretical variogram to any given type of process. We have to take account both empirical knowledge of the phenomenon and the shape of the experimental variogram obtained. As this is the nugget effect, one might think that it is not appropriate for data such as rainfall.

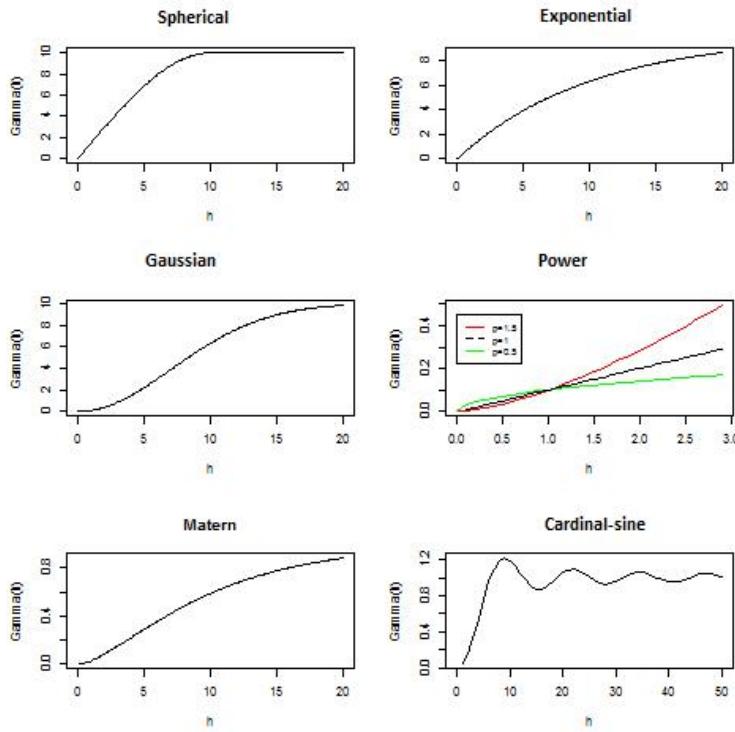


Figure 5.10 – Four examples of theoretical variograms

### 5.3.3 Variogram fitting

A functional form appropriate to the experimental variogram must then be found. An important first step is to obtain a smoothed representation of the variogram, which can be obtained using the `variog` function of `geoR`. As always, this smoothing depends on the choice of window. The smoothed representation is sometimes considered adequate to visually estimate the variogram. Many geostatisticians criticise this approach as too empirical, but this step can give some indications, particularly on behaviour at the origin.

Figure 5.11 shows three examples of fitting the experimental variogram to rainfall data in Switzerland, using a spherical variogram, an exponential variogram without nugget and an exponential variogram with nugget. These adjustments are made using the `lines.variomodel` function of `geoR` (Ribeiro Jr et al. 2006). This is the first visual approach. We can see that while the exponential variogram with nugget apparently fits better to the data than the variogram without nugget, the introduction of this very short distance effect has no physical justification for rainfall.

---

```
library(geoR)
vario.ex<- variog(sic.100,option="bin")
vario.sphe<- (variofit(vario.ex,cov.model= "spher",
ini.cov.pars=c(15000,200)))
par(mfrow=c(2,2), mar=c(3,3,1,1), mgp =c (2,1,0))
plot(vario.ex,main="Spherical")
lines.variomodel(cov.model="sphe",cov.pars=c(15000,100),
nug=0,max.dist=350)
plot(vario.ex,main="Exponential")
lines.variomodel(cov.model="exp",cov.pars=c(15000,100),
nug=0,max.dist=350)
```

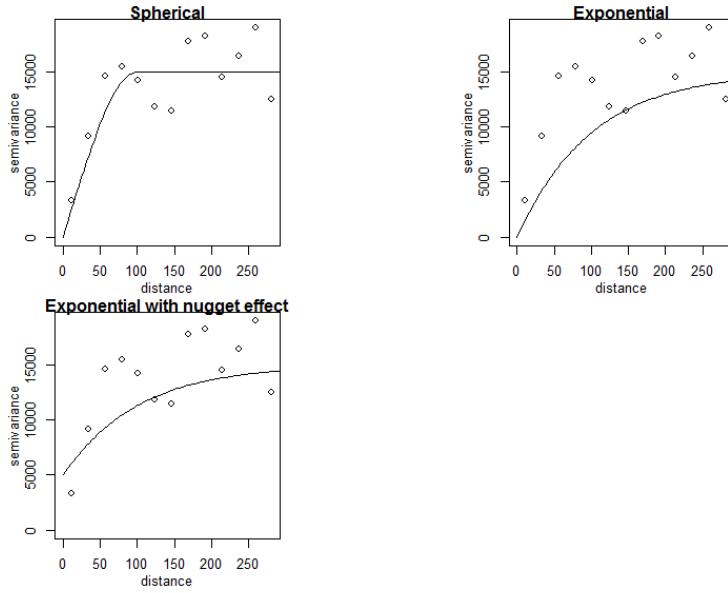


Figure 5.11 – Three examples of experimental variogram adjustment

**Source:** Swiss rainfall from the geoR package

```
plot(vario.ex, main="Exponential with nugget effect")
lines.variomodel(cov.model="exp", cov.pars=c(10000,100),
nug=5000, max.dist=350)
plot(vario.ex, main="Exponential with nugget effect")
lines.variomodel(cov.model="matern", cov.pars=c(10000,100),
nug=0, max.dist=350, kappa=0.5)
```

The choice is therefore a compromise as noted by Waller et al. 2004: "Even if a particular model is deemed better for a particular dataset using a statistical adjustment method, it may not necessarily be the best choice. For example, the Gaussian model is often selected using an automatic adjustment criterion, but this provides smoothing that often appears unrealistic. Ultimately, the final choice of model should reflect both the result of the procedure for adjusting the statistical model and a consistent interpretation with scientific understanding of the process being studied."

Many methods are suggested to fit the variogram — methods based on ordinary or weighted least squares, methods based on likelihood, as well as Bayesian methods. In *geoR*, the functions used are *variofit* (least squares) and *likfit* (maximum likelihood). The methods are quite technical and would require significant developments. Refer to Ribeiro Jr et al. 2006 for an illustration of these methods using simulated data. The R code is supplied in the article.

### Fitting by Ordinary Least Squares (OLS)

We look for the vector of the function's parameters that minimises a simple objective function, the sum of squares of the distances between the value of the experimental semi-variogram and the value of the theoretical variogram.

$$\hat{\theta}_{MCO} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^k (\hat{\gamma}(h_i) - \gamma(h_i; \theta))^2 \quad (5.21)$$

### Fitting by Weighted Least Squares (WLS)

Ordinary least squares do not take into account the number of point pairs involved in the calculation of each point of the experimental variogram, unlike the weighted least squares.

$$\hat{\theta}_{MCP} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^k \frac{\#N(h_i)}{\gamma(h_i; \theta)^2} (\hat{\gamma}(h_i) - \gamma(h_i; \theta))^2 \quad (5.22)$$

### Fitting by Generalised Least Squares (GLS)

Other authors have proposed generalised least squares to take heteroscedasticity into account.

$$\hat{\theta}_{MCG} = \underset{\theta \in \Theta}{\operatorname{argmin}} (\hat{\gamma}_n - \gamma(\theta))^T \operatorname{Cov}(\gamma_n)^{-1} (\hat{\gamma}_n - \gamma(\theta)) \quad (5.23)$$

where  $\gamma$  is the vector  $(\gamma_1, \gamma_2, \dots, \gamma_K)$ .

### Fitting by maximum likelihood (ML)

The parameters of the model are estimated by calculating the likelihood. In the non-Gaussian case, the estimates are not robust. The calculations are tedious, and this method must only be used for small samples. In addition, this method requires second order stationarity and cannot be applied to unbounded variograms. In the latter case, the weighted least squares must be used.

## 5.4 Ordinary kriging

### 5.4.1 Principle

The term kriging was coined by Georges Matheron, and refers to the pioneering work by South African engineer Danie Krige. Kriging is a very powerful interpolation method. The examples provided here are very basic. Applications to mining or geological research provide many examples where we want to estimate volumes and not just simple interpolations. We will not discuss *simple* kriging here, which assumes the mean value is known, but *ordinary* kriging, which forms the highlight of geostatistics. In ordinary kriging, the mean value is unknown. Simple uses of it can be found in interpolating temperatures (Joly et al. 2009) or in air quality studies (Lloyd et al. 2004). Assume that  $Z(\cdot)$  is intrinsically stationary, that its variogram  $\gamma(h)$  is known but its mean  $m$  is unknown. We have a data set  $Z = [Z(s_1), \dots, Z(s_i), \dots, Z(s_N)]^T$ . We want to predict the value of  $Z(\cdot)$  as an unobserved point and calculate  $Z(s_0)$ . The ordinary kriging estimator will be defined as a linear combination of observations.

$$Z_{OK}(s_0) = \sum_{i=1}^N \lambda_i Z(s_i) \quad (5.24)$$

The weighting values  $\lambda_i$  are not calculated using a distance function but by using the semi-variogram and two statistical criteria. This way, there is no bias and minimisation of the mean squared prediction error. The absence of bias brings us to Equation 5.25:

$$E [\hat{Z}_{OK}(s_0)] = E [Z(s_0)] = m \quad (5.25)$$

$$E [\hat{Z}_{OK}(s_0)] = \sum_{i=1}^N \lambda_i E[Z(s_i)] = \sum_{i=1}^N \lambda_i m \quad \rightarrow \quad \sum_{i=1}^N \lambda_i = 1$$

Therefore, using the Lagrange multipliers method, we will minimise  $E [\hat{Z}_{OK}(s_0) - Z(s_0)]^2$  under the constraint  $\sum_{i=1}^N \lambda_i = 1$ .

Box 5.4.1 shows how to introduce the variogram and end up with the kriging equations.

$$\begin{aligned} \sum_{j=1}^N \lambda_j \gamma(s_i - s_j) + m &= \gamma(s_0 - s_i) \\ \sum_{i=1}^N \lambda_i &= 1 \end{aligned} \quad (5.26)$$

The value of  $\hat{Z}_{OK}(s_0)$  is determined by points that depend on the correlation between the estimation point and the observation points, but also correlations between the observation points. These kriging equations are generally best written in matrix form:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ m \end{bmatrix} = \begin{bmatrix} \gamma(s_1 - s_1) & \dots & \gamma(s_1 - s_N) & 1 \\ \gamma(s_2 - s_1) & \dots & \gamma(s_2 - s_N) & 1 \\ \vdots & \ddots & \ddots & \vdots \\ \gamma(s_N - s_1) & \dots & \gamma(s_N - s_N) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(s_0 - s_1) \\ \gamma(s_0 - s_2) \\ \vdots \\ \vdots \\ \gamma(s_0 - s_N) \end{bmatrix} \quad (5.27)$$

or in a more condensed form:

$$\lambda_0 = \Gamma_{ij}^{-1} \gamma_0. \quad (5.28)$$

Matrix  $\Gamma$  does not depend on the estimation point, and does not therefore have to be recalculated every time. The values of all  $\gamma(s_i - s_j)$  and  $\gamma(s_0 - s_i)$  are calculated using values from the estimated variogram. The mean squared prediction error  $r$ , known as *kriging variance*, is equal to  $\lambda_0^t \gamma_0$ . In short, kriging provides an unbiased estimator of minimum variance that is also an exact interpolator since, for every known point, it returns an estimated value equal to the observed value.

**Box 5.4.1 — Estimate of kriging equations.** Using the Lagrange multipliers method, we minimise:

$$E [\hat{Z}_{OK}(s_0) - Z(s_0)]^2 \text{ under the constraint } \sum_{i=1}^N \lambda_i = 1.$$

We'll look for  $\lambda_1, \dots, \lambda_N$  and multiplier  $m$  that enable the constraint to be introduced. The objective function is therefore written:

$$E \left[ \left( \sum_{i=1}^N \lambda_i Z(s_i) - Z(s_0) \right)^2 \right] - 2m \left( \sum_{i=1}^N \lambda_i - 1 \right). \quad (5.29)$$

Due to the constraint, we can write:

$$\left[ \sum_{i=1}^N \lambda_i Z(s_i) - Z(s_0) \right]^2 = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \left[ (Z(s_i) - Z(s_j))^2 \right] + \sum_{i=1}^N \left[ (Z(s_i) - Z(s_0))^2 \right]. \quad (5.30)$$

By taking the expected value of the expressions, we have:

$$E \left[ \sum_{i=1}^N \lambda_i Z(s_i) - Z(s_0) \right]^2 = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j E \left[ (Z(s_i) - Z(s_j))^2 \right] + \sum_{i=1}^N \lambda_i E \left[ (Z(s_i) - Z(s_0))^2 \right]. \quad (5.31)$$

This expression brings out the variogram and we can rewrite the constraint as:

$$-\sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \gamma(s_i - s_j) + 2 \sum_{i=1}^N \gamma(s_0 - s_i) - 2m \left( \sum_{i=1}^N \lambda_i - 1 \right). \quad (5.32)$$

We will minimise this expression by deriving in relation to  $\lambda_1, \dots, \lambda_N$  and  $m$ , which leads to the kriging equations:

$$\begin{aligned} \sum_{j=1}^N \lambda_j \gamma(s_i - s_j) + m &= \gamma(s_0 - s_i) \\ \sum_{i=1}^N \lambda_i &= 1. \end{aligned} \quad (5.33)$$

## 5.4.2 Application to rainfall data

### Raw data

An initial kriging was carried out from raw data using a spherical model for the variogram. Figure 5.12 shows the spherical variogram used, after estimating the parameters using maximum likelihood compared to the experimental variogram.

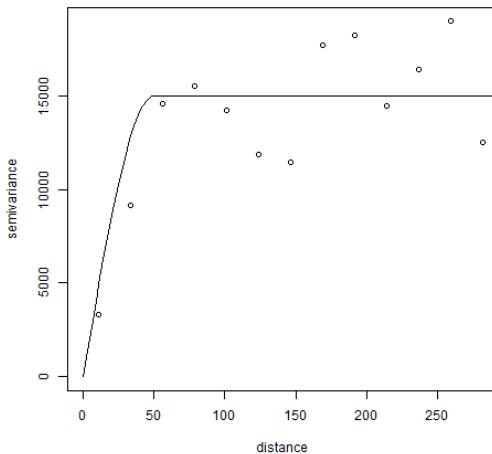


Figure 5.12 – Spherical variogram for raw data

**Source:** Swiss rainfall from the geoR package

---

```
library(geoR)
vario.ex<- variog(sic.100, bin.cloud=TRUE)
plot(vario.ex,main="")
lines.variomodel(cov.model="spher",cov.pars=c(15000,50),
nug=0,max.dist=300)
```

---

This way, we can calculate kriged values on a grid, as well as the kriging variances represented in Figure 5.13. The values are represented according to graphical semiology conventions, the warm colours corresponding to the high values.

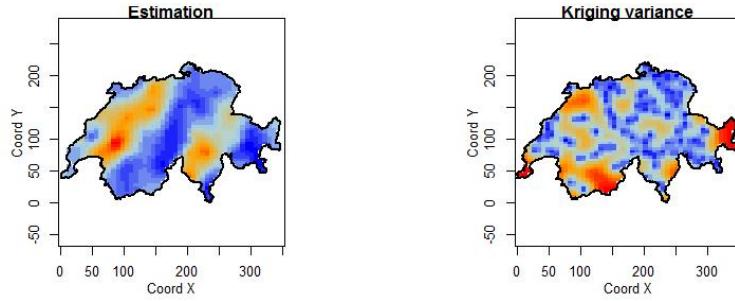


Figure 5.13 – Kriging estimates and variances for raw data

**Source:** Swiss rainfall from the geoR package

---

```
library(geoR)
pred.grid <- expand.grid(seq(0,350, l=51),seq (0,220, l=51))
rgb.palette <- colorRampPalette(c("blue", "lightblue",
"orange", "red"),space = "rgb")
kc<- krige.conv(sic.100, loc = pred.grid,
krige=krige.control(cov.model="spherical",cov.pars=c(15000,50)))
image(kc, loc = pred.grid,col =rgb.palette(20) ,xlab="Coord X",
ylab="Coord Y",borders=sic.borders,main="Estimation")
image(kc, krige.var,loc = pred.grid,col=rgb.palette(20),
xlab="Coord X",ylab="Coord Y",borders=sic.borders,
main="Kriging variance")
```

---

The estimate was made using the 100 points of the sample. We can:

- confirm that kriging is unbiased on the points observed;
- measure differences between estimated values and observed values.

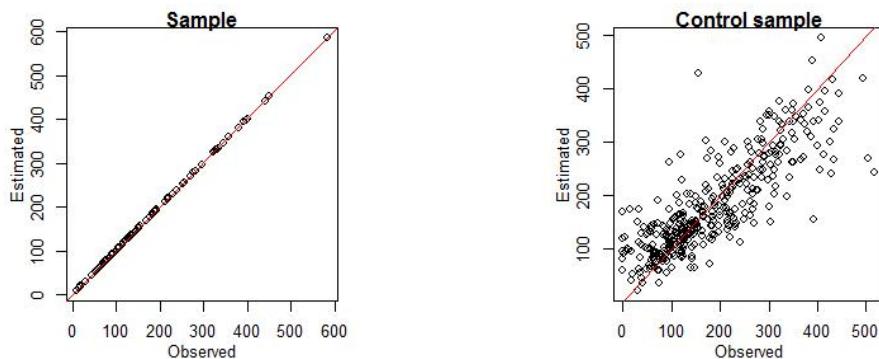


Figure 5.14 – Estimated and observed values

**Source:** Swiss rainfall from the geoR package

---

```
library(geoR)
kc1<- krige.conv(sic.100, loc = sic.100$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47)))
kc2<- krige.conv(sic.100, loc = sic.367$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47)))
plot(sic.100$data,kc1$predict,xlab="Observed",ylab="Estimated",
```

---

---

```
main="Control sample")
abline(a=0,b=1,col="red")
plot(sic.367$data,kc2$predict,,xlab="Observed",ylab="Estimated",
main="Control")
abline(a=0,b=1,col="red")
```

---

### Transformed data

The histogram in Figure 5.3 indicates that the distribution of rainfall data deviated from a Gaussian distribution. A first possibility, classical in statistics, is to modify variables. In fact, it may be better to work with data that obey a normal law. This is not absolutely necessary in the kriging model, but the linearity assumption in kriging is only really effective when the data are Gaussian. Classical transformations are logarithmic transformations, or more generally Box-Cox transformations. The `boxcoxfit` function of *geoR* suggests a coefficient close to 0.5. Exploratory data are shown in Figure 5.15.

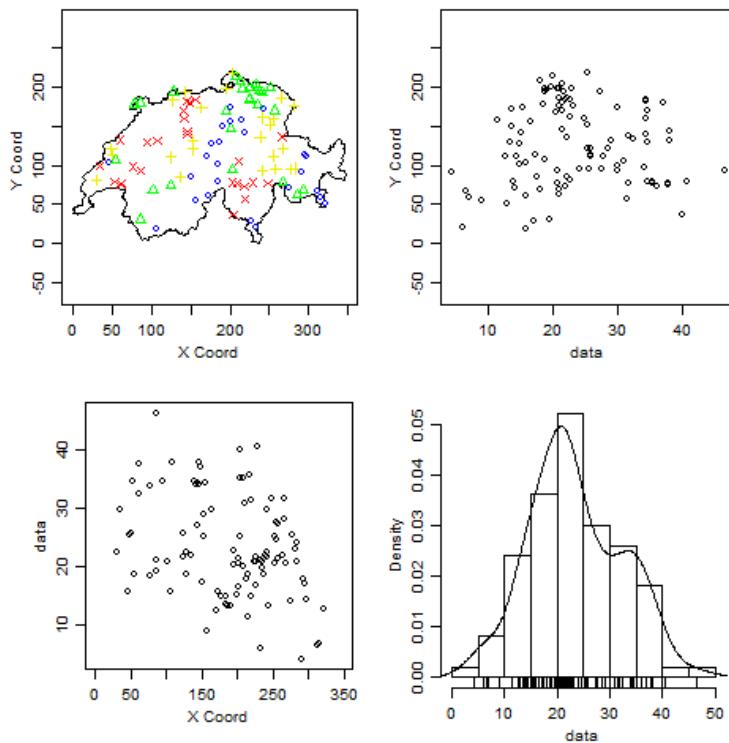


Figure 5.15 – Exploratory data, after Box-Cox transformation

**Source:** Swiss rainfall from the *geoR* package

---

```
library(geoR)
plot.geodata(sic.100,bor=sic.borders,lambda=0.5)
```

---

Rainfall data have frequently been studied. Ribeiro Jr et al. 2004 recommends the same transformation of variables. For the variogram, they suggest using a Matern model for which  $K = 1$ . The parameters of the model are determined using maximum likelihood. The experimental and theoretical variograms for transformed data are shown in Figure 5.16.

---

```
library(geoR)
```

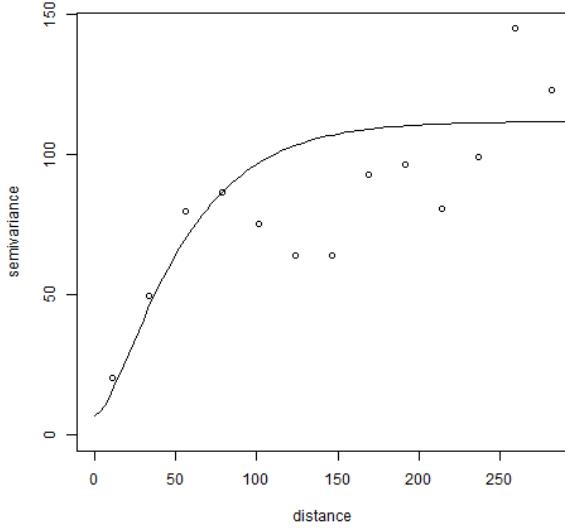


Figure 5.16 – Experimental variogram and theoretical variogram after transformation  
**Source:** Swiss rainfall from the geoR package

---

```
vario.ext<- variog(sic.100,option="bin",lambda=0.5)
plot(vario.ext)
lines.variomodel(cov.m = "mat",cov .p =c (105, 36), nug = 6.9,
max.dist = 300,kappa = 1, lty = 1)
```

---

As for raw data, we can provide mapping of kriging estimates and values (Figure 5.17).

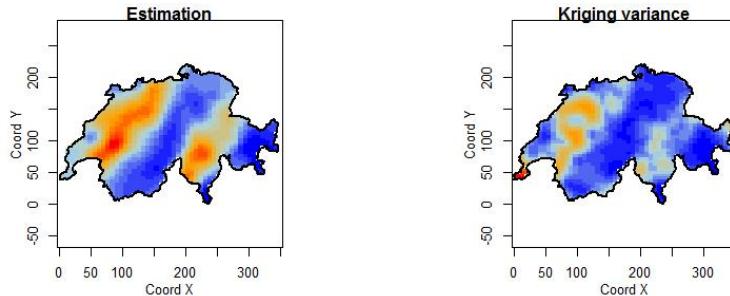


Figure 5.17 – Kriging estimates and variance of rainfall in Switzerland after transformation  
**Source:** Swiss rainfall from the geoR package

---

```
library(geoR)
kct<- krige.conv(sic.100, loc = pred.grid,
krige.control(cov.model="matern",cov.pars=c(105, 36),
kappa=1,nugget=6.9,lambda=0.5))
pred.grid <- expand.grid(seq(0,350, l=51),seq (0,220, l=51))
rgb.palette <- colorRampPalette(c("blue", "lightblue",
"orange", "red"),space = "rgb")
image(kct, loc = pred.grid,col =rgb.palette(20) , xlab="Coord X",
ylab="Coord Y",borders=sic.borders,main="Estimation")
```

---

---

```
image(kct, krige.var, loc = pred.grid, col = rgb.palette(20) ,
xlab="Coord X", ylab="Coord Y", borders=sic.borders,
main="Kriging variance")
```

---

The estimated values and observed values can be compared (5.18).

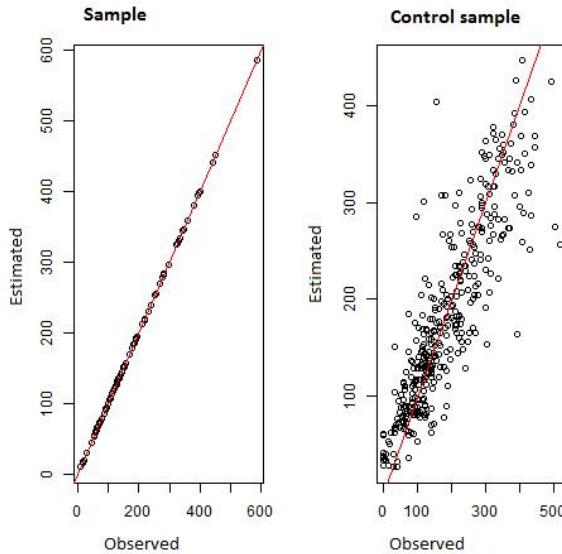


Figure 5.18 – Estimated and observed values

**Source:** Swiss rainfall from the geoR package

---

```
library(geoR)
kct1<- krige.conv(sic.100, loc = sic.100$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47),
kappa=1,nugget=6.9,lambda=0.5))
kct2<- krige.conv(sic.100, loc = sic.367$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47),
kappa=1,nugget=6.9,lambda=0.5))
plot(sic.100$data,kct1$predict,xlab="Observed",ylab="Estimated",
main="Sample")
abline(a=0,b=1,col="red")
plot(sic.367$data,kct2$predict,,xlab="Observed",ylab="Estimated",
main="Control sample")
abline(a=0,b=1,col="red")
```

---

If we take the square root of the mean quadratic deviation as a criterion,

$$RMSE(y) = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

we see an improvement in prediction if we use the transformed data modelled by the Matern model. In the case of raw data, the value of RMSE is 62.3, while for transformed data it is 55.2.

## 5.5 Support and change of support

The question of analysis scales is of primary importance in spatial analysis. Geographers, especially Openshaw, use the term *Modifiable areal unit problem* (MAUP), discussed in Chapter 1: "Descriptive spatial analysis". The MAUP can be summarised as overlaying a zoning effect and an aggregation effect. This can be illustrated simply using surface data, by showing the variability of results obtained depending on the territorial breakdown used.

In geostatistics, we talk about *change of support problem* (COSP). This problem is actually more general than MAUP, since it refers to size, shape and directions. The COSP in geostatistics originates from very practical concerns linked to mining research, the origins of which can be found in Krige's work, which was then expanded on by Matheron. Krige's articles are contemporary with those by Yule and Kendall in classical statistics, which anticipate those by Openshaw in geography. From a practical viewpoint, geostatisticians quickly realised that it was more important to predict a value over a large block than on one point, even though the first prediction derived from the second.

The support can be a point, a larger or smaller block, or a meeting of points in a given geometric configuration. There is a relationship between the values taken on different volumes. In the context of additive variables, for a volume  $V$  partitioned in units  $v_i$  of same support  $v$  ( $V$  being a multiple of  $v$ ), we have:

$$z(V) = \frac{1}{n} \sum_{i=1}^n z(v_i) \quad (5.34)$$

or if  $v$  is unique:

$$z(V) = \frac{1}{V} \int_V z(x) dx. \quad (5.35)$$

$Z(V)$  is qualified as a *regularised* variable, because it increases statistical regularity. In fact, it is a specific form of regularised variable, the more general form being a *Z convolute* (Chiles et al. 2009).

### 5.5.1 Empirical dispersion variance and Krige additivity relationships

**Definition 5.5.1 — Empirical dispersion variation.**

$$s^2(v|V) = \frac{1}{n} \sum_{i=1}^n [z(v_i) - z(V)]^2 \quad (5.36)$$

If  $V$  is included in a broader domain called  $D$ , we can demonstrate relationship 5.37, called the **Krige additivity relationship**:

$$s^2(v|D) = s^2(v|V) + s^2(V|D). \quad (5.37)$$

Armstrong 1998 gives an educational example from the yields in millet on 16 blocks of 2m side, divided into 64 parcels of 1m side. The average is 201 in the case of blocks as in the case of parcels. The variance of the dispersion of the blocks in the field is 16.64, that of the parcels 27.59, which means that the dispersion of the parcels in the blocks is 10.85. (see 5.1).

### 5.5.2 Variogram of the regularised variable

Variance by block can be defined based on information about single data points (covariance function).

$$\text{Var}[Z(V)] = \bar{C}(V, V) = \frac{1}{|V|^2} \int_V \int_V C(x - y) dx dy \quad (5.38)$$

735	325	45	140	125	175	167	485
540	420	260	128	20	30	105	70
450	200	337	190	95	260	245	278
180	250	380	405	250	80	515	605
124	120	430	175	230	120	460	260
40	135	240	35	190	135	160	170
75	95	20	35	32	95	20	450
200	35	100	59	2	45	58	90

505	143	88	207
270	328	171	411
102	220	154	263
101	54	44	155

Table 5.1 – Observed values on parcels (top) and blocks (bottom)

**Source :** Armstrong 1998.

where  $C$  designates the covariance function for single data points and  $\bar{C}$  designates covariance of the data by block. Covariance is written:

$$\text{Cov}[Z(V), Z(V')] = \bar{C}(V, V') = \frac{1}{|V||V'|} \int_V \int_{V'} C(x-y) dx dy. \quad (5.39)$$

Covariance function  $C_V$  is defined by introducing  $V_h$  which is the translation of support  $V$  by vector  $h$ :

$$C_V(h) = \text{Cov}[Z(V), Z(V_h)] = \bar{C}(V, V_h). \quad (5.40)$$

From the unique variogram we can also deduce the variogram of the regularised variable:

$$\gamma_V(h) = \bar{\gamma}(V, V_h) - \bar{\gamma}(V, V) \quad (5.41)$$

with  $\gamma(V, V) = \frac{1}{|V|^2} \int_V \int_V \gamma(x-y) dx dy$  and  $\bar{\gamma}(V, V_h) = \frac{1}{|V|^2} \int_V \int_{V_h} \gamma(x-y) dx dy$  where  $\gamma$  is the variogram calculated from unique observations.

We can then show that  $\gamma_V(h) \sim \gamma(h) - \bar{\gamma}(V, V)$ , leading to the graph in Figure 5.19.

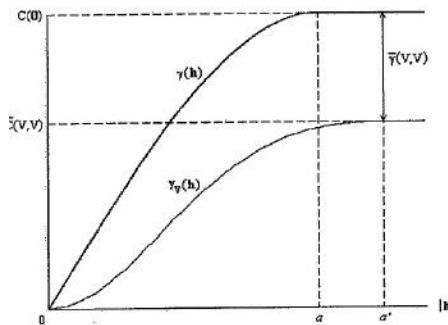


Figure 5.19 – Switch from unique variogram to regularised variogram

To switch from unique to regularised variogram, the same type of theoretical model is retained by correcting the sill and the range.

### 5.5.3 Block kriging

We can determine kriging equations, in the logic of ordinary kriging. From  $E[Z(x)] = m$ , we deduce that  $E[Z_V] = m$ .

As for ordinary kriging, we are looking for an estimator that is a linear combination of the observations collected.

$$Z_V = \sum_{i=1}^n \lambda_i Z(x_i) \quad (5.42)$$

The approach is the same as that shown in box 5.4.1 (minimisation under constraint) and we get the following kriging equations:

$$\begin{aligned} \sum_{j=1}^n \lambda_j \gamma(x_i, x_j) + \mu &= \bar{\gamma}(x_i, V) \quad \text{pour } i = 1, 2, \dots, n \\ \sum_{i=1}^n \lambda_i &= 1 \end{aligned} \quad (5.43)$$

where  $\bar{\gamma}(x_i, V) = \frac{1}{|V|} \int_V \gamma(x_i - x) dx$ .

## 5.6 Extensions

Ordinary kriging is the basic method of geostatistics. But in the end, it is only part of it, and there have been many developments, particularly at the Ecole des Mines in Fontainebleau.

Intrinsic stationarity assumptions remain fairly restrictive, in particular the constancy of the mean. Many methods have been developed to introduce less constraining assumptions, or to use supplementary information. In this paragraph, a number of modifications and variants will be presented, mainly from Swiss rainfall data, and another frequently-used dataset on the contents of various minerals in a meander of the River Meuse.

### 5.6.1 Cokriging

Geostatistics developed multivariate methods long ago. One of them is Cokriging, which will consider several variables. It was defined by Waller et al. 2004 as follows:

**Definition 5.6.1 — cokriging.** "Cokriging is an extension of kriging to the case of two or more spatial variables. It was originally developed as a technique for improving the prediction of a variable for which only a few samples could be taken, by using its spatial correlation with other more easily measured variables. Cokriging differs from kriging with external drift in that the explanatory variables are no longer assumed to be fixed variables that indicate the nature of a trend in the primary variable, but are themselves spatial random variables with expected values and variograms.".

A cross co-variogram can be defined:

$$\gamma_{ZY}(h) = \frac{1}{2p(h)} \sum_{i=1}^{p(h)} (z(s_i) - z(s_i + h)) (y(s_i) - y(s_i + h)) \quad (5.44)$$

with  $p(h) = \text{Card} \{(s_i, s_j) | |s_i - s_j| \approx h\}$

Just like kriging, there will be several versions for cokriging. We will only discuss ordinary cokriging. We will restrict ourselves to the case where only one auxiliary variable is introduced, which will be called  $Y$ . The estimator we calculate takes the form:

$$Z(s_0) = \sum_{i=1}^{n_Z} \lambda_i Z(s_i) + \sum_{i=1}^{n_Y} \alpha_i Y(s_i) \quad (5.45)$$

with bias-free constraints:

$$\begin{aligned} \sum_{i=1}^{n_Z} \lambda_i &= 1 \\ \sum_{i=1}^{n_Y} \alpha_i &= 0. \end{aligned} \quad (5.46)$$

In matrix form, the cokriging equations are written:

$$\begin{bmatrix} \Gamma_{ZZ} & \Gamma_{ZY} & 1 & 0 \\ \Gamma_{YZ} & \Gamma_{YY} & 0 & 1 \\ 1' & 0' & 0 & 0 \\ 0 & 1' & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \alpha \\ \mu_Z \\ \mu_Y \end{bmatrix} = \begin{bmatrix} \gamma_{ZZ} \\ \gamma_{YZ} \\ 1 \\ 0 \end{bmatrix} \quad (5.47)$$

This method will be illustrated using River *Meuse* data supplied in the *sp* package and particularly studied by Pebesma (Pebesma 2001) and Rossiter (Rossiter 2017). The latter author provides R programs on his website.

### ■ Example 5.1 Analysis of River *Meuse* data by Cokriging

The River *Meuse* data provide localised measurements of lead, zinc and cadmium contents, but also other variables such as altitude and organic matter content of the soil. The *sp* package contains a table called *Meuse*, which can be loaded with the `data(meuse)` function. It also provides a 40x40 m grid – *meuse.grid* – and boundaries of the "département" – *meuse.riv*.

The data comprise 155 observations on 15x15 m supports, over the upper 20 cm of alluvial soils on the right bank of the River Meuse. Associated data provide the geographical coordinates of the observations, their altitude, and concentrations of cadmium, copper, lead, zinc and organic matter. There is also the distance to the River Meuse and the flooding frequency. In the example provided by Rossiter 2017, the lead content is studied (after logarithmic transformation). The organic matter content that will be used as covariate.

The variogram analysis for Cokriging is based on studying the two simple variograms and the cross-variogram.

The same exercise is performed using the logarithm of the zinc content as covariate. The figure below shows the kriging results for the lead content, by ordinary kriging and by cokriging, using the two variables mentioned above (zinc content and organic matter). Figure 5.22 shows the kriged values (left column) and the residual values (right column). It successively shows ordinary kriging, cokriging with organic matter as co-variable, and cokriging with the zinc content.

The code required for this processing is quite long. The link below provides access to the R programme made available by Rossiter (Rossiter 2007)

[http://www.css.cornell.edu/faculty/dgr2/teach/R/ck\\_plotfns.R](http://www.css.cornell.edu/faculty/dgr2/teach/R/ck_plotfns.R)

If we compare the three models using RMSE, we find the following results:

- 0.166 for ordinary kriging;

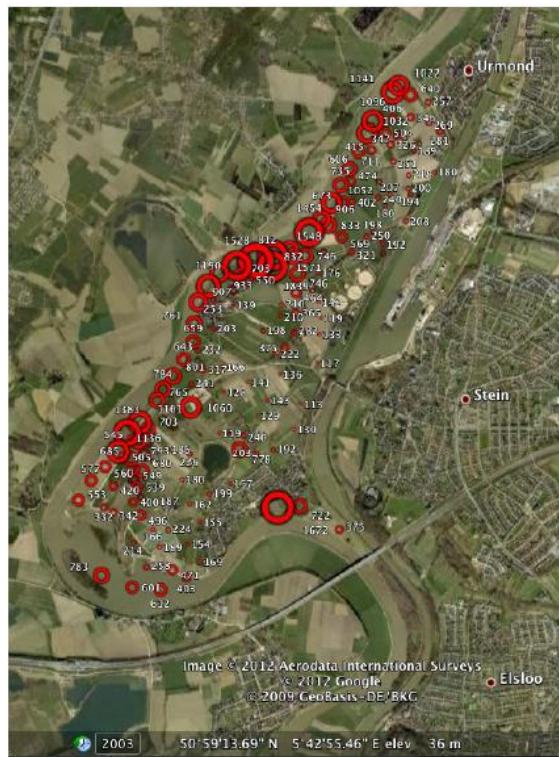


Figure 5.20 – Geography and sample locations

**Source:** *Meuse data from the sp package*

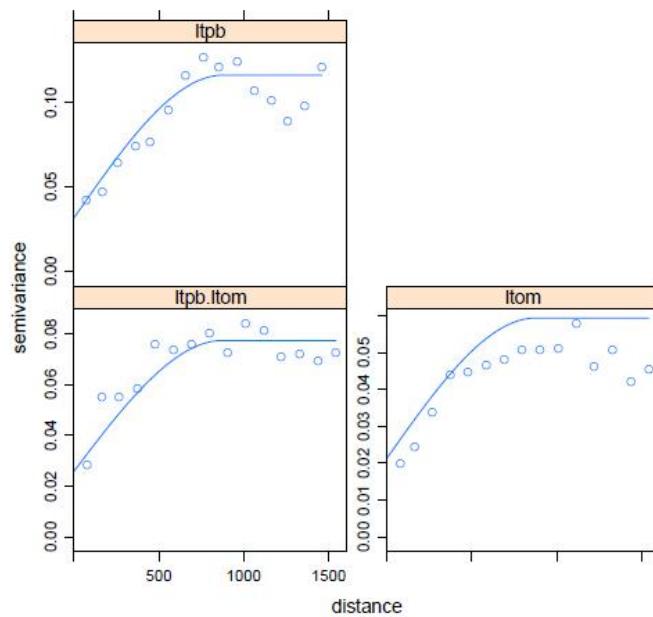


Figure 5.21 – Simple and cross variograms

**Source:** *Meuse data from the sp package*

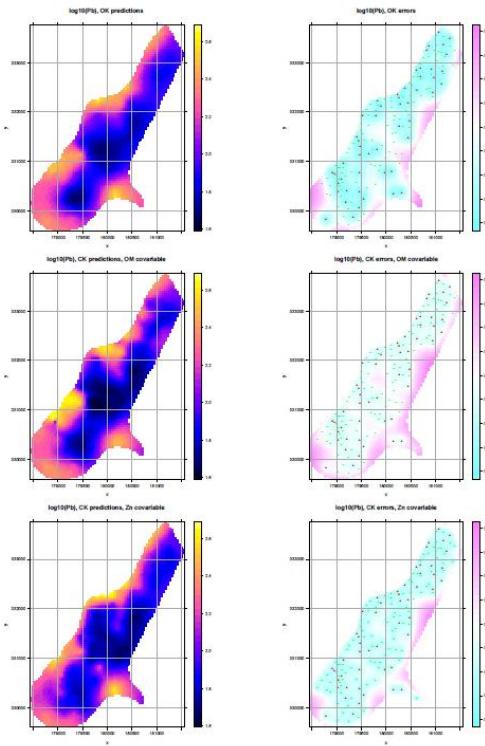


Figure 5.22 – Kriging and Cokriging estimate and variance

**Source:** *Meuse data from the sp package*

- 0.226 for cokriging with organic matter content;
- 0.078 for cokriging with zinc content.

Cokriging involving zinc content therefore improves the performance of ordinary kriging. This is not true for cokriging with organic matter content. ■

### 5.6.2 Universal kriging

In many cases, the average value is not constant, and ordinary kriging cannot be used. This applies when deterministic relationships are observed between the value of the variable and its position in space. The regionalized variable can then be written:

$$Z(s) = m(s) + Y(s) \quad (5.48)$$

where  $m(s)$  represents the deterministic component.

■ **Example 5.2 — Analysis of River Meuse data by cokriging.** The River Meuse data provide two variables likely to build a deterministic component — distance to the river and flooding frequency (Figure 5.23). They are different from the co-variables used previously in cokriging.

On his website, Rossiter provides examples used to compare the predictions from normal kriging and from two universal kriging models.

If we compare the three models using RMSE, we find the following results:

- 0.173 for ordinary kriging;
- 0.141 for the model with flooding frequency;
- 0.145 for the model with flooding frequency and distance to the river.

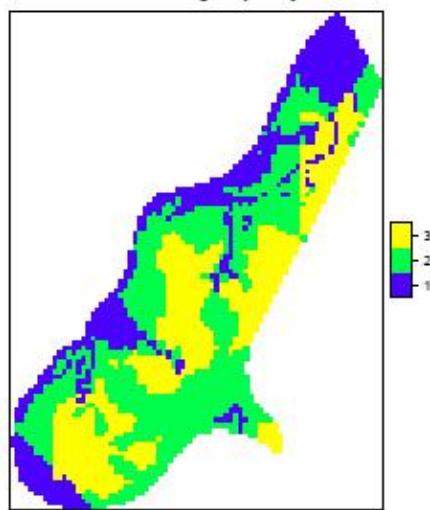


Figure 5.23 – Flooding frequency

**Source:** Meuse data from the *sp* package

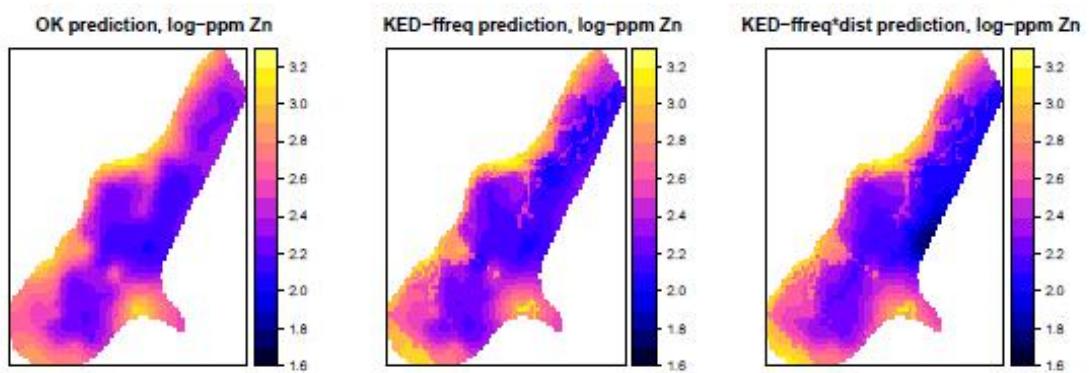


Figure 5.24 – Three estimates

**Note:** ordinary kriging (left), flooding frequency (middle), flooding frequency and distance to the river (right)

**Source:** Meuse data from the *sp* package

Introducing flooding frequency improves predictions, but introducing the distance also degrades it (probably due to correlation of the two variables). The R code used is available at the address below. It is a good example to illustrate and extend what has been discussed in this chapter.

[http://www.css.cornell.edu/faculty/dgr2/teach/R/gs\\_short\\_ex.pdf](http://www.css.cornell.edu/faculty/dgr2/teach/R/gs_short_ex.pdf)

■

## 5.7 Combined models with variogram

Other uses can be found for tools developed in geostatistics. When performing regressions on spatial data, the residuals are frequently spatially autocorrelated. This correlation can be highlighted by the Moran spatial autocorrelation indicator (see Chapter 3: "Spatial autocorrelation indices"). This autocorrelation can be taken into account using spatial econometric models (see Chapter 6: "Space econometrics: current models") or geographically-weighted regression (see Chapter 9: "Geographically-weighted regression").

When the data are suitable, the variogram can also be used to study the spatial structure of residuals in linear models. Examples are more frequently described in books dealing with ecology or epidemiology. Illustrations can be found in general spatial statistical handbooks such as Waller et al. 2004 or Schabenberger et al. 2017, as well as books dealing with ecology, such as Plant 2012 or Zuur et al. 2009, the last two providing examples of implementation in R.

■ **Example 5.3 — Analysis of the spatial structure of residuals with a variogram.** An example of use on ecological data is provided by Zuur et al. 2009<sup>2</sup>. The example comes from data collected over forests in the Raifa section of the Voljsko-Kamsky state natural biosphere.

The interest variable is a 'boreality' index (*Bor*) defined as the share of specifically boreal species compared to the total number of species on a site. There are also explanatory variables provided by satellite images:

1. standardised vegetation difference index;
2. temperature;
3. moisture index;
4. greening index.

Due to the strong co-linearity between these variables, only moisture was used to explain the boreality index. The variance analysis performed using ordinary least squares provides the following results:

Variable	Estimated value	standard deviation
Constant	27.63	0.981
Wet	429.609	27.45

Table 5.2 – Least squares estimate

The addresses of the sites are used to provide an exploratory vision of the spatialisation of residues of the OLS model.

---

```
library(sp)
library(nlme)
Boreality<-read.table("C:/jmf/Boreality.txt",header=TRUE)
B.lm <- lm(boreal ~ Wet, data = Boreality)
```

---

2. [https://github.com/James-Thorson/2016\\_class\\_CMR/tree/master/Other%20material/Zuur%20et%20al.%202007/ZuurDataMixedModelling](https://github.com/James-Thorson/2016_class_CMR/tree/master/Other%20material/Zuur%20et%20al.%202007/ZuurDataMixedModelling)

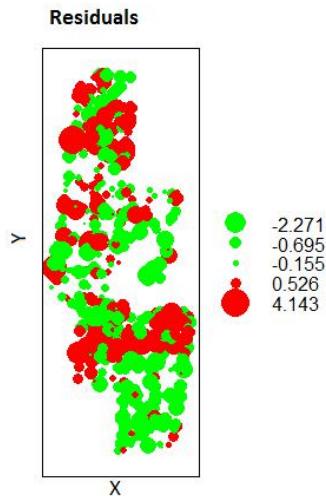


Figure 5.25 – Residuals

**Source:** Meuse data from the *sp* package

```
E <- rstandard(B.lm)
graphic <- data.frame(E, Boreality$x, Boreality$y)
library(sp)
coordinates(graphic) <- c("Boreality.x", "Boreality.y")
bubble(graphic, "E", col = c ("green","red"),main = "Residuals",
xlab = "X", ylab = "Y")
```

The OLS model does not allow the introduction of spatial structure on residues. It can only be introduced in a generalised linear model. It will be estimated using the *gls* function of the R *nlme* package. This package contains a function to estimate the variogram.

The experimental variogram is shown below (Figure 5.26)

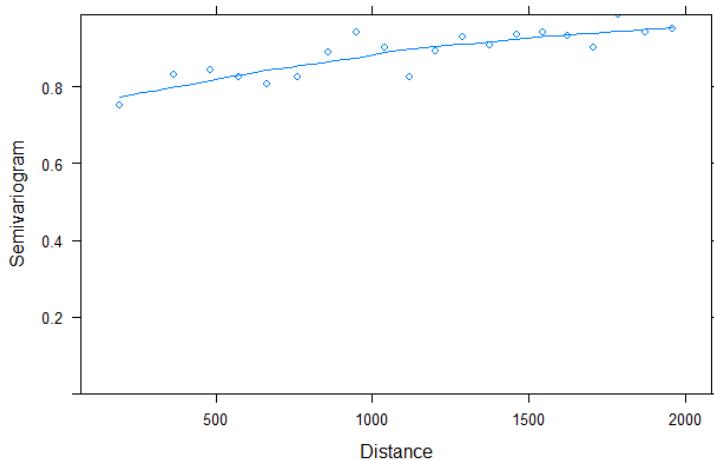


Figure 5.26 – Variogram of residuals

**Source:** Meuse data from the *sp* package

---

```
mod<-gls(boreal~Wet, data=bor)
summary(mod)
plot(Variogram(mod, form=~x+y, maxdist=10000), xlim=c(0,10000))
```

---

The initial estimate from the model, without the introduction of the spatial structure, gives the same results as OLS. The update command is used to introduce a spatial structure using a variogram analysis and to re-estimate the model.

Table 5.3 shows the result of the comparison between OLS and models using spherical, Gaussian and exponential variograms.

Variogram	AIC	Likelihood	L	Significance
OLS	3855	-1924		
Spherical	3859	-1924	1	1
Gaussian	3750	-1870	109	<0.001
Exponential	3740	-1865	119	<0.001

Table 5.3 – AIC and L criteria based on variogram type

The AIC and L criteria show that the model can be improved by using a spherical variogram to model the residues. The regression results with the exponential variogram are shown in Table 5.4.

Variable	Estimated value	standard deviation
constant	18.099	2.333
Wet	180.247	34.932

Table 5.4 – Estimates

---

```
f1 <- formula(boreal ~ Wet)
B1.gls <- gls(f1, data = Boreality)
Vario.gls <- Variogram(B1.gls, form =~ x + y, robust = TRUE,
maxDist = 2000, resType = "pearson")
B1A <- gls(f1, correlation = corSpher(form =~ x + y, nugget = TRUE),
data = Boreality)
B1B <- gls(f1, correlation = corLin(form =~ x + y, nugget = TRUE),
data = Boreality)
B1C <- gls(f1, correlation = corRatio(form =~ x + y, nugget = TRUE),
data = Boreality)
B1D <- gls(f1, correlation = corGaus(form =~ x + y, nugget = TRUE),
data = Boreality)
B1E <- gls(f1, correlation = corExp(form =~ x + y, nugget = TRUE),
data = Boreality)
AIC(B1.gls, B1A, B1B, B1C, B1D, B1E)
B1 <- lm(f1, data = Boreality)
anova(B1.gls, B1A)
anova(B1.gls, B1D)
anova(B1.gls, B1E)
summary(B1E)
```

---

The parameters of the model are significant. The influence of moisture is less pronounced when spatial autocorrelation is introduced into the model. ■

## **Conclusion**

The first chapter of this handbook presents the three main areas of spatial statistics appropriate to the analysis of continuous, surface or single data points. Geostatistics, used for continuous data, is less directly linked to the work of public statistics. Nonetheless, it seemed useful to provide a quick description in the handbook. From an educational viewpoint, geostatistical methods illustrate particularly well how considering spatial autocorrelation (through the variogram) makes it possible to improve the estimators. From a more operational viewpoint, without going into the complexity of mining research work, geostatistics *using kriging methods* is useful for modelling simpler continuous data (*e.g.* climate data). The Fontainebleau Ecole des Mines, which has played a crucial part in developing of these methods, has used quite unusual language for statisticians, but many discussions have been held since Cressie's work to link the different approaches. The classic book by Chilès and Delfiner is a good example of this (Chiles et al. 2009). In the healthcare field, significant work has involved geostatistical methods to model epidemiological data, particularly that by Diggle (Diggle et al. 2003), who is better known for his work on *ad hoc* methods. Finally, we can only recommend that statisticians who may use models in the future should read the exploratory article by the founder of geostatistics (Matheron 1978).

## Appendices

### Mathematical reminders

The expressions of theoretical variograms, in particular Matern's variogram, use quite unusual mathematical expressions, in particular the expression below (Chiles et al. 2009).

#### The Gamma function

$$\Gamma(x) = \int_0^\infty e^{-u} u^{x-1} du$$

For whole number values:

$$\Gamma(n+1) = n!$$

#### The Bessel functions

The Bessel function of the first kind is as follows:

$$J_v(x) = \left(\frac{x}{2}\right)^v \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(v+k+1)} \left(\frac{x}{2}\right)^{2k}$$

The modified Bessel function of the first kind is as follows:

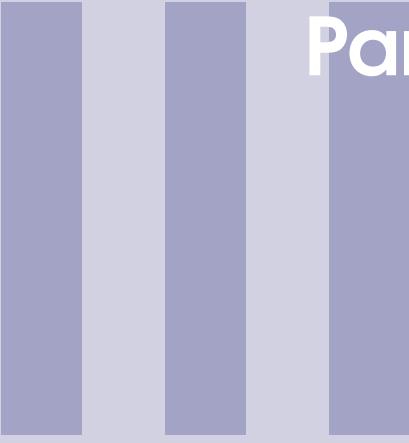
$$I_v(x) = \left(\frac{x}{2}\right)^v \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(v+k+1)} \left(\frac{x}{2}\right)^{2k}$$

The modified Bessel function of the second kind is defined from the previous type

$$K_v(x) = \frac{\pi}{2} \frac{I_{-v}(x) + I_v(x)}{\sin \pi v}$$

## References - Chapter 5

- Armstrong, Margaret (1998). *Basic linear geostatistics*. Springer Science & Business Media.
- Chauvet, Pierre (2008). *Aide-mémoire de géostatistique linéaire*. Presses des MINES.
- Chiles, Jean-Paul and Pierre Delfiner (2009). *Geostatistics: modeling spatial uncertainty*. Vol. 497. John Wiley & Sons.
- Diggle, Peter J, Paulo J Ribeiro Jr, and Ole F Christensen (2003). « An introduction to model-based geostatistics ». *Spatial statistics and computational methods*. Springer, pp. 43–86.
- Joly, Daniel et al. (2009). « Interpolation par régressions locales: application aux précipitations en France ». *L'Espace géographique* 38.2, pp. 157–170.
- Lloyd, Christopher D and Peter M Atkinson (2004). « Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data ». *International Journal of Applied Earth Observation and Geoinformation* 5.4, pp. 293–305.
- Matheron, Georges et al. (1965). *Les variables régionalisées et leur estimation*. Masson et Cie.
- Matheron, Georges (1978). *Estimer et choisir: essai sur la pratique des probabilités*. Ecole nationale supérieure des mines de Paris.
- Pebesma, Edzer J (2001). « Gstat user's manual ». *Dept. of Physical Geography, Utrecht University, Utrecht, The Netherlands*.
- Plant, Richard E (2012). *Spatial data analysis in ecology and agriculture using R*. cRc Press.
- Ribeiro Jr, Paulo J and Peter J Diggle (2006). « geoR: Package for Geostatistical Data Analysis An illustrative session ». *Artificial Intelligence* 1, pp. 1–24.
- Ribeiro Jr, Paulo Justiniano and Peter J Diggle (2004). « Model Based Geostatistics ». *Springer Series in Statistics*.
- Rossiter, David G (2017). « An introduction to geostatistics with R/gstat Version 3.7, 12-May-2017. »
- Rossiter, DG (2007). « Co-kriging with the gstat package of the R environment for statistical computing ». Web: [http://www.itc.nl/rossiter/teach/R/R\\_ck.pdf](http://www.itc.nl/rossiter/teach/R/R_ck.pdf).
- Schabenberger, Oliver and Carol A Gotway (2017). *Statistical methods for spatial data analysis*. CRC press.
- Waller, Lance A and Carol A Gotway (2004). *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons.
- Zuur, AF et al. (2009). « Mixed effects models and extensions in ecology with R. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors ». *New York, NY: Spring Science and Business Media*.



# Part 3: Taking spatial effects into account

6	Spatial econometrics - common models	149
7	Spatial econometrics on panel data .	179
8	Spatial smoothing .....	205
9	Geographically Weighted Regression	231
10	Spatial sampling .....	255
11	Spatial econometrics on survey data	277
12	Small area estimation and spatial correlation .....	305



# 6. Spatial econometrics - common models

JEAN-MICHEL FLOCH

INSEE

RONAN LE SAOUT

ENSAI

---

<b>6.1</b>	<b>What are the benefits of taking spatial, organisational or social proximity into account?</b>	<b>151</b>
6.1.1	The economic reasons .....	151
6.1.2	Econometric reasons .....	152
<b>6.2</b>	<b>Autocorrelation, heterogeneity and weightings: a review of key points in spatial statistics</b>	<b>152</b>
6.2.1	The nature of spatial effects in regression models .....	152
6.2.2	The weight matrix .....	153
6.2.3	Exploratory methods .....	153
<b>6.3</b>	<b>Estimating a spatial econometrics model</b>	<b>154</b>
6.3.1	The galaxy of spatial econometrics models .....	154
6.3.2	Statistical criteria for model selection .....	155
6.3.3	When interpreting results, beware of feedback effects .....	157
<b>6.4</b>	<b>Econometric limits and challenges</b>	<b>160</b>
6.4.1	What to do with missing data? .....	160
6.4.2	Choosing the weight matrix .....	160
6.4.3	What if the phenomenon is spatially heterogeneous? .....	161
6.4.4	The risk of “ecological” errors .....	162
<b>6.5</b>	<b>Practical application under R</b>	<b>163</b>
6.5.1	Mapping and testing .....	163
6.5.2	Estimation and model selection .....	164
6.5.3	Interpreting the results .....	169
6.5.4	Other spatial modelling .....	171

---

## 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<sup>1</sup> 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 remains 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<sup>2</sup> 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  $I^* = \frac{I - E(I)}{\sqrt{V(I)}}$  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  $\beta$  a parameter to be estimated,  $\varepsilon \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$  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<sup>3</sup> :

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

with parameters  $\beta$  for exogenous explanatory variables,  $\rho$  for the endogenous interaction effect (of dimension 1) referred to as spatial autoregressive,  $\theta$  for exogenous interaction effects (of dimension equal to the number of exogenous variables  $K$ ) and  $\lambda$  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.*  $\beta$ ,  $\rho$ ,  $\theta$ , and  $\lambda$  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  $W_\rho$  reflecting the autoregressive parameter and  $W_\lambda$  reflecting the spatial autocorrelation. Slade 2005 defines two separate neighbourhood matrices to study price effects in industrial economy:  $W_\rho$  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  $\rho$ ,  $\theta$  and  $\lambda$ . 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  $\beta_0$  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 invertibility of matrices  $I - \rho W$  and  $I - \lambda W$ , and the identification of models. It can be noted that the usual patterns of contiguity 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  $\beta$  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  $\varepsilon \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ ,

they are based on an estimate by maximum likelihood of the models and the related statistical tests<sup>4</sup>. 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.

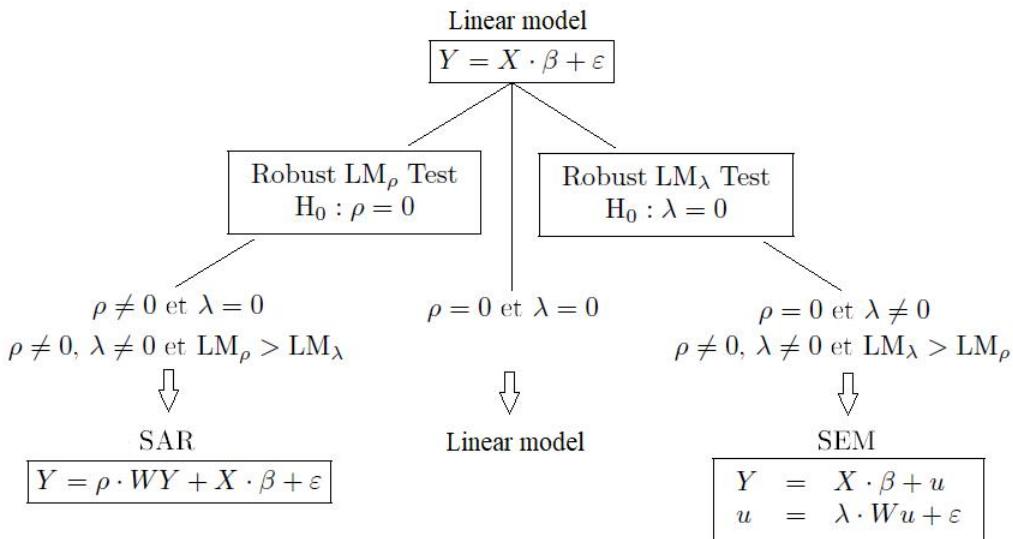


Figure 6.1 – The *bottom-up* approach

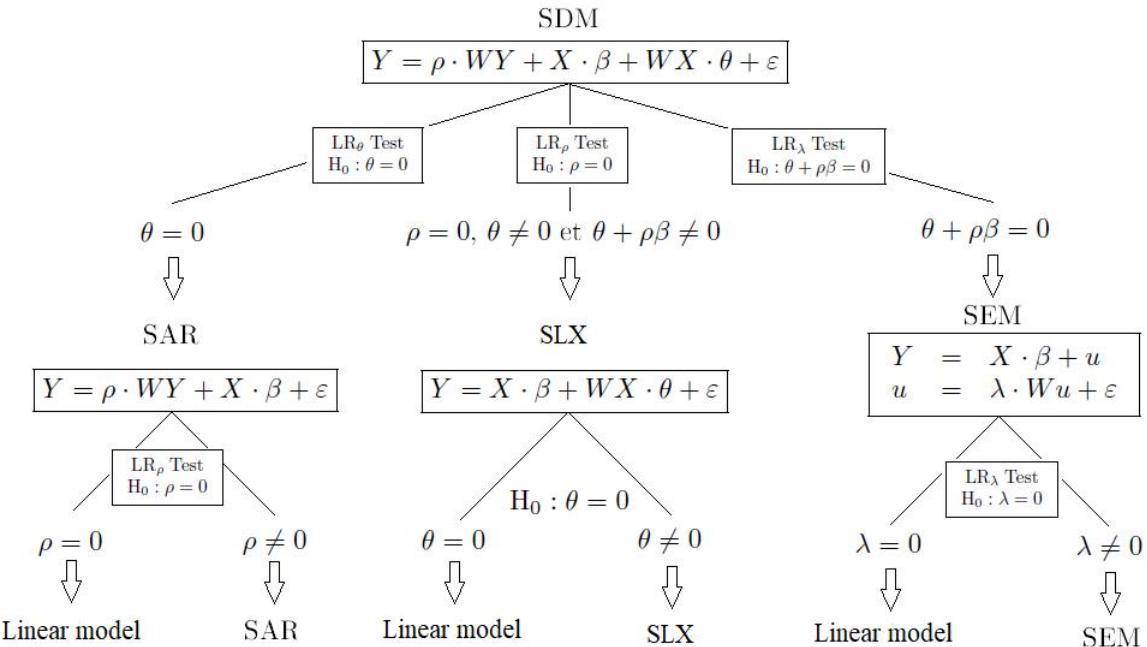
**Source :** Florax et al. 2003

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 ( $LM_\rho$  and  $LM_\lambda$ )<sup>5</sup> 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.*  $\rho \neq 0$  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<sup>6</sup>, 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.

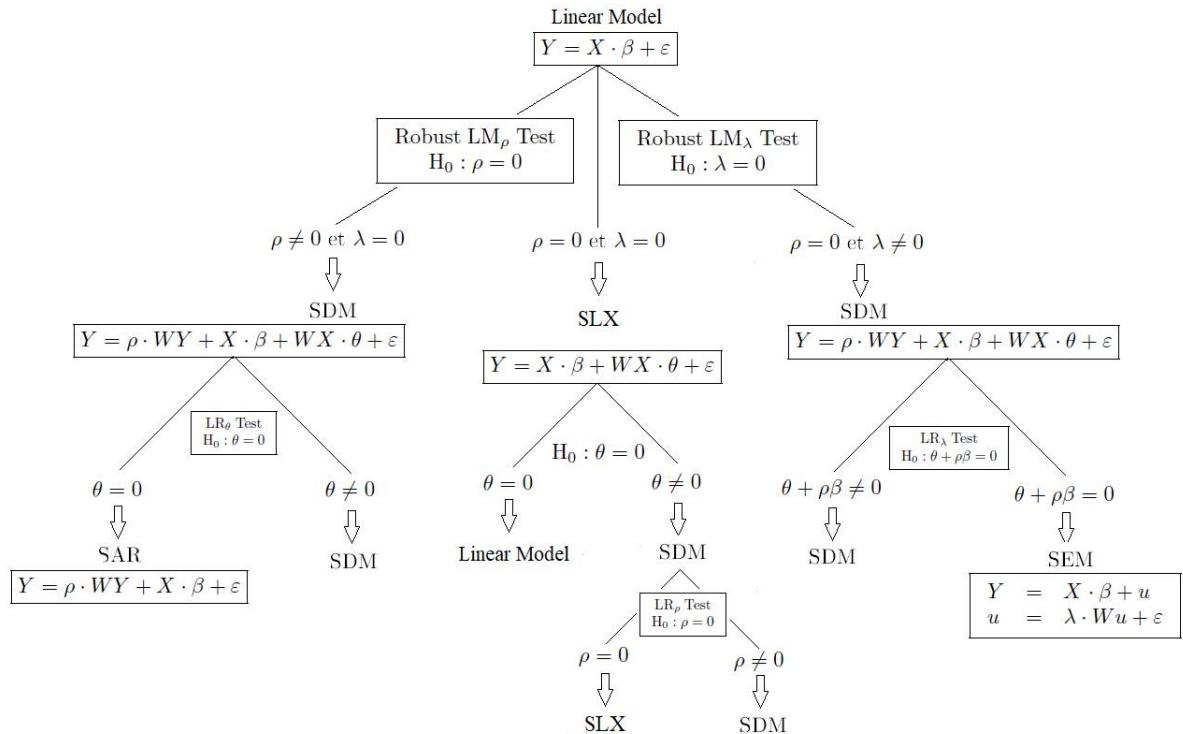


Figure 6.3 – Approach proposed by Elhorst 2010 for choosing a spatial econometric model

**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:

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

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

The predicted value is therefore  $\hat{y} = (1 - \hat{\rho}W)^{-1}X\hat{\beta}$ <sup>7</sup> 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  $\beta_r$  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<sup>8</sup>. 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  $\beta$  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  $\beta_r$ , while the indirect effect is  $\theta_r$ .

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)<sup>9</sup>.

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  $(1 - \rho W)^{-1} = (I_n + \rho W + \rho^2 W^2 + \dots)$  shows that the first feedback term (which dominates the other higher order terms) is proportional to  $\rho^2$ . The analysis of effects by neighbourhood order (distinguishing the direct effect, the effect of neighbours, neighbours of neighbours, etc.) is also elaborated upon by Lesage 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  $\Phi$  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 consists in estimating the average marginal effect  $\beta_r \cdot \varphi(\beta X)$ .

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 ( $\frac{1}{n} [\sum_j \sum_k S_r(W)_{kj} - \text{trace}(S_r)]$ ). 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<sup>10</sup>, 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  $\varepsilon$  from the models presented in this introduction. Kelejian et al. 2007/Kelejian 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 (Brenson 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  $1 \text{ km}^2$  or in  $10000 \text{ km}^2$ . 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  $\beta$  represents the individual effect and  $\gamma$  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  $\beta$  and  $\gamma$ . The model is heteroskedastic because  $\text{V}(\bar{\varepsilon}_k) = \sigma^2/n_k$  in the case of initial disturbances independent and identically distributed of variance  $\sigma^2$ .

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 contiguity 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 contiguous 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.

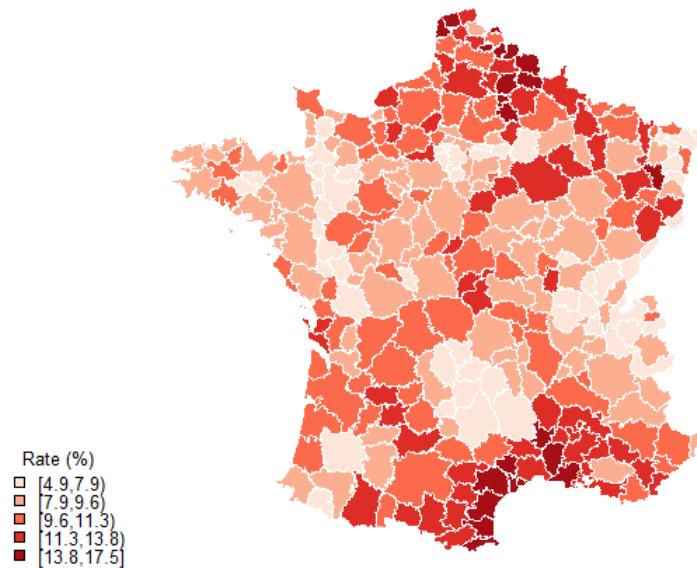
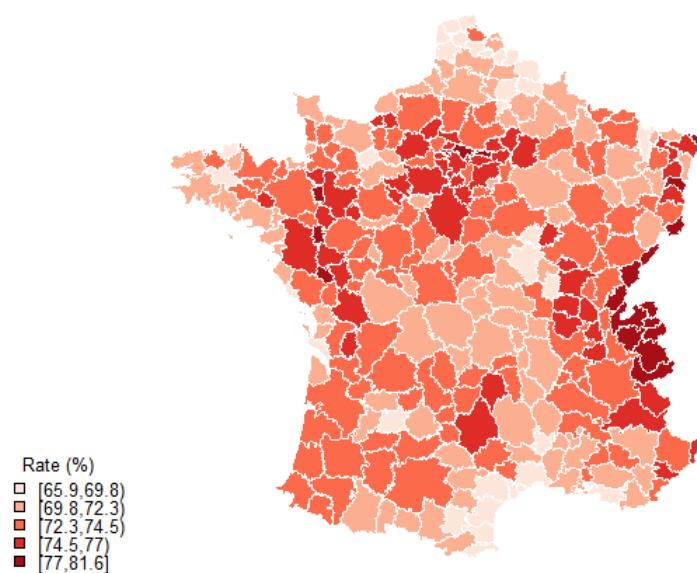
**Unemployment rate (2013)****Labour force participation rate (2011)**

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

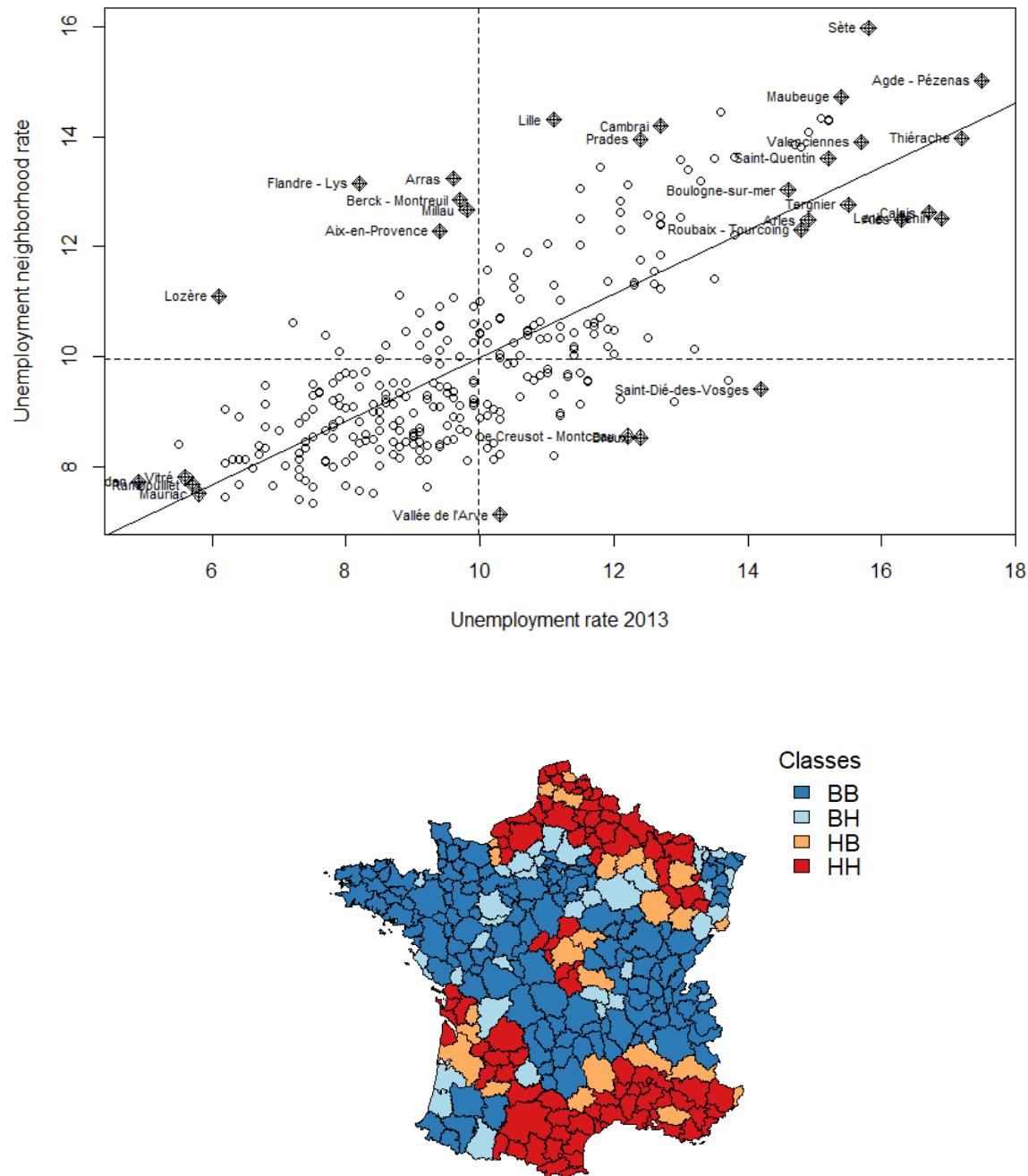


Figure 6.5 – Moran unemployment rate graph and associated map

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, contiguity 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*** (0.039)	-0.498*** (0.041)	-0.437*** (0.038)	-0.472*** (0.042)	-0.499*** (0.041)	-0.470*** (0.050)	-0.486*** (0.041)	-0.473*** (0.042)
% Low educated working-age adults	0.186*** (0.026)	0.184*** (0.027)	0.138*** (0.022)	0.182*** (0.027)	0.179*** (0.026)	0.179*** (0.033)	0.181*** (0.027)	0.183*** (0.028)
% Working-age adults 15-30 y.o.	0.138*** (0.043)	0.196*** (0.045)	0.087** (0.037)	0.209*** (0.046)	0.180*** (0.045)	0.205*** (0.055)	0.197*** (0.045)	0.211*** (0.047)
% Industrial employment	-0.062*** (0.012)	-0.018 (0.012)	-0.036*** (0.010)	-0.015 (0.012)	-0.021* (0.012)	-0.022 (0.014)	-0.024** (0.012)	-0.014 (0.012)
% Public employment	-0.068*** (0.019)	-0.044*** (0.016)	-0.063*** (0.016)	-0.042** (0.016)	-0.048*** (0.017)	-0.044** (0.019)	-0.049*** (0.017)	-0.041** (0.016)
$\hat{\rho}$			0.519*** (0.049)	0.629*** (0.064)	0.205* (0.109)			0.689*** (0.120)
$\hat{\lambda}$		0.747*** (0.051)			0.616*** (0.096)		0.651*** (0.063)	-0.137 (0.257)
$\hat{\theta}$ , Participation rate				0.157* (0.083)		-0.300*** (0.082)	-0.277*** (0.105)	0.205* (0.111)
$\hat{\theta}$ , % Low educated working-age adults				-0.135*** (0.045)		-0.027 (0.052)	-0.021 (0.066)	-0.145*** (0.046)
$\hat{\theta}$ , % Working-age adults 15-30 y.o.				-0.140 * (0.072)		-0.041 (0.085)	-0.003 (0.115)	-0.153** (0.072)
$\hat{\theta}$ , % Industrial employment				-0.044** (0.020)		-0.118*** (0.023)	-0.073** (0.029)	-0.038* (0.023)
$\hat{\theta}$ , % Public employment				-0.024 (0.037)		-0.084* (0.043)	-0.070 (0.052)	-0.018 (0.037)
Intercept	51.653*** (3.635)	39.729*** (3.685)	34.470*** (3.407)	27.456*** (6.766)	38.427*** (3.901)	66.077*** (6.514)	63.650*** (10.213)	23.530*** (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*** (0.039)	-0.518*** (0.040)	-0.517*** (0.040)	-0.530*** (0.040)	-0.507*** (0.040)	-0.498*** (0.041)	-0.515*** (0.041)
% Low educated working-age adults	0.186*** (0.026)	0.188*** (0.026)	0.204*** (0.026)	0.185*** (0.026)	0.181*** (0.026)	0.184*** (0.027)	0.184*** (0.026)
% Working-age adults 15-30 y.o.	0.138*** (0.043)	0.179*** (0.045)	0.195*** (0.044)	0.201*** (0.045)	0.198*** (0.046)	0.196*** (0.045)	0.139*** (0.044)
% Industrial employment	-0.062*** (0.012)	-0.023* (0.012)	-0.027** (0.012)	-0.023* (0.012)	-0.024** (0.012)	-0.018 (0.012)	-0.026** (0.012)
% Public employment	-0.068*** (0.019)	-0.042** (0.017)	-0.039** (0.017)	-0.047*** (0.017)	-0.048*** (0.017)	-0.044*** (0.016)	-0.050*** (0.016)
$\hat{\lambda}$		0.687*** (0.050)	0.506*** (0.047)	0.681*** (0.051)	0.763*** (0.053)	0.747*** (0.051)	0.700*** (0.044)
Intercept	51.653*** (3.635)	41.535*** (3.681)	40.672*** (3.643)	42.166*** (3.639)	40.685*** (3.644)	39.729*** (3.685)	42.414*** (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 [-0.700,-0.545]	-0.509 [-0.588,-0.435]	-0.510 [-0.589,-0.434]	-0.529 [-0.611,-0.451]	-0.505 [-0.583,-0.422]	-0.490 [-0.574,-0.409]	-0.508 [-0.588,-0.429]
% Low educated working-age adults	0.186 [0.136,0.237]	0.178 [0.122,0.232]	0.208 [0.154,0.261]	0.183 [0.132,0.235]	0.177 [0.125,0.230]	0.180 [0.122,0.230]	0.178 [0.129,0.232]
% Working-age adults 15-30 y.o.	0.138 [0.054,0.223]	0.194 [0.102,0.288]	0.223 [0.135,0.312]	0.213 [0.123,0.309]	0.212 [0.119,0.306]	0.207 [0.119,0.299]	0.184 [0.092,0.279]
% Industrial employment	-0.062 [-0.087,-0.038]	-0.026 [-0.048,-0.003]	-0.032 [-0.053,-0.008]	-0.027 [-0.051,-0.005]	-0.027 [-0.050,-0.005]	-0.022 [-0.045,0.001]	-0.033 [-0.055,-0.011]
% Public employment	-0.068 [-0.106,-0.030]	-0.045 [-0.078,-0.010]	-0.048 [-0.081,-0.011]	-0.052 [-0.084,-0.017]	-0.051 [-0.083,-0.018]	-0.049 [-0.081,-0.014]	-0.052 [-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 [-0.587,-0.091]	-0.200 [-0.337,-0.068]	-0.241 [-0.488,0.007]	-0.306 [-0.700,0.030]	-0.357 [-0.658,-0.073]	-0.351 [-0.638,-0.107]
% Low educated working-age adults	-0.015 [-0.161,0.142]	-0.059 [-0.146,0.032]	-0.032 [-0.205,0.124]	-0.050 [-0.291,0.158]	-0.053 [-0.254,0.137]	-0.079 [-0.251,0.085]
% Working-age adults 15-30 y.o.	-0.016 [-0.321,0.249]	-0.079 [-0.214,0.058]	-0.082 [-0.334,0.174]	0.016 [-0.321,0.390]	-0.023 [-0.352,0.301]	0.047 [-0.230,0.332]
% Industrial employment	-0.130 [-0.208,-0.055]	-0.064 [-0.105,-0.022]	-0.100 [-0.170,-0.030]	-0.135 [-0.244,-0.041]	-0.136 [-0.229,-0.059]	-0.111 [-0.187,-0.043]
% Public employment	-0.120 [-0.274,0.017]	-0.078 [-0.140,-0.011]	-0.113 [-0.257,0.031]	-0.098 [-0.345,0.132]	-0.130 [-0.335,0.046]	-0.037 [-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

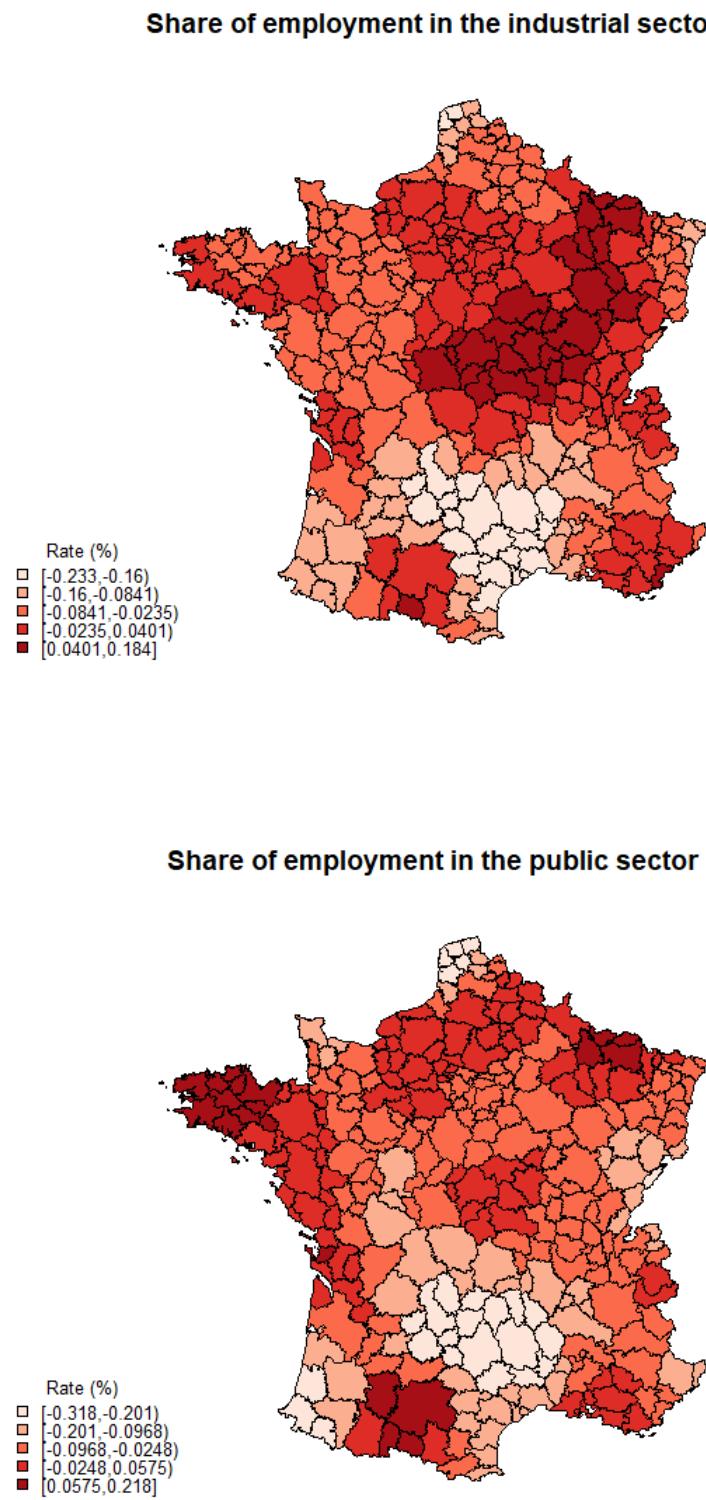


Figure 6.6 – Distribution of local parameters

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*** (0.042)	-0.485*** (0.042)	-0.513*** (0.041)	-0.494*** (0.041)	-0.472*** (0.042)	-0.485*** (0.042)
% Low educated working-age adults	0.180*** (0.027)	0.215*** (0.028)	0.186*** (0.028)	0.179*** (0.027)	0.182*** (0.027)	0.184*** (0.028)
% Working-age adults,15-30 y.o.	0.196*** (0.047)	0.232*** (0.046)	0.219*** (0.047)	0.211*** (0.048)	0.209*** (0.046)	0.181*** (0.047)
% Industrial employment	-0.016 (0.012)	-0.024** (0.012)	-0.020* (0.012)	-0.022* (0.012)	-0.015 (0.012)	-0.026** (0.012)
% Public employment	-0.037** (0.017)	-0.038** (0.017)	-0.044*** (0.017)	-0.048*** (0.017)	-0.042** (0.016)	-0.050* (0.016)
$\hat{\rho}$	0.601*** (0.057)	0.448*** (0.050)	0.606*** (0.057)	0.647*** (0.068)	0.629*** (0.064)	0.609*** (0.051)
$\hat{\theta}$ , Participation rate	0.153** (0.075)	0.094 (0.057)	0.209*** (0.072)	0.207** (0.087)	0.157* (0.083)	0.149** (0.075)
$\hat{\theta}$ , % Low educated working-age adults	-0.114*** (0.040)	-0.133*** (0.034)	-0.126*** (0.041)	-0.134*** (0.047)	-0.135*** (0.045)	-0.145*** (0.040)
$\hat{\theta}$ , % Working-age adults 15-30 y.o.	-0.124* (0.069)	-0.153*** (0.053)	-0.167*** (0.065)	-0.131* (0.078)	-0.140* (0.072)	-0.090 (0.068)
$\hat{\theta}$ , % Industrial employment	-0.046** (0.021)	-0.029** (0.015)	-0.030 (0.019)	-0.035 (0.022)	-0.044** (0.020)	-0.031* (0.018)
$\hat{\theta}$ , % Public employment	-0.029 (0.033)	-0.031 (0.022)	-0.020 (0.031)	-0.005 (0.043)	-0.024 (0.037)	0.015 (0.031)
Intercept	28.582*** (6.184)	33.848*** (4.814)	26.710*** (5.844)	24.504*** (7.372)	27.456*** (6.766)	27.662*** (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](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). « Endogénéité 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-Ltrémy, 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*.



# 7. Spatial econometrics on panel data

**BOUAYAD AGHA SALIMA**

*GAINS (TEPP) and CREST*

*Le Mans Université*

**LE GALLO JULIE**

*CESAER, AgroSup Dijon, INRA,*

*Université de Bourgogne Franche-Comté, F-21000 Dijon*

**VÉDRINE LIONEL**

*CESAER, AgroSup Dijon, INRA,*

*Université de Bourgogne Franche-Comté, F-21000 Dijon*

---

<b>7.1</b>	<b>Specifications</b>	<b>180</b>
7.1.1	Standard model: modelling individual specific effects . . . . .	180
7.1.2	Spatial effects in panel data models . . . . .	181
7.1.3	Interpretation of coefficients in the presence of a spatial autoregressive term . . . . .	185
<b>7.2</b>	<b>Estimation methods</b>	<b>186</b>
7.2.1	Fixed effects model . . . . .	186
7.2.2	Random effects model . . . . .	187
<b>7.3</b>	<b>Specification tests</b>	<b>189</b>
7.3.1	Choosing between fixed and random effects . . . . .	189
7.3.2	Specification tests for spatial effects . . . . .	189
<b>7.4</b>	<b>Empirical application</b>	<b>190</b>
7.4.1	The model . . . . .	190
7.4.2	Data and spatial weights matrix . . . . .	193
7.4.3	The results . . . . .	194
<b>7.5</b>	<b>Extensions</b>	<b>198</b>
7.5.1	Dynamic spatial models . . . . .	198
7.5.2	Multidimensional spatial models . . . . .	199
7.5.3	Panel models with common factors . . . . .	200

---

## Abstract

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

 Prior reading of Chapter 6: “Spatial econometrics: common models” is recommended.

## Introduction

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

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

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

## 7.1 Specifications

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

### 7.1.1 Standard model: modelling individual specific effects

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

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

$$y_{it} = x_{it}\beta + z_i\alpha + \varepsilon_{it} \quad (7.1)$$

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

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

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

$$y_{it} = x_{it}\beta + \alpha + \varepsilon_{it} \quad (7.2)$$

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

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

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it} \quad (7.3)$$

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

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

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

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

Unlike the fixed-effect model, individual effects are no longer parameters to be estimated, but realisations of a random variable. This model is therefore appropriate if individual specificities are linked to random causes. It is also preferable to the fixed effects model when the individuals in the sample are drawn from a larger population and the objective of the empirical study is to generalise to the population the results obtained. This model offers the advantage of providing more accurate estimates than those derived from the fixed effects model. It is usually estimated using the Generalised Least Squares (GLS) method.

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

### 7.1.2 Spatial effects in panel data models

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

### Spatial effects in pooled data models

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

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

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

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

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

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

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

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

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

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

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

### Spatial effects in fixed effects models

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

### Spatial Error Model-Random Effect

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

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

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

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

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

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

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

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

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

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

Lastly, a more general specification as suggested by Baltagi et al. 2007<sup>2</sup>:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

## 7.2 Estimation methods

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

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

### 7.2.1 Fixed effects model

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

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

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

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

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

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

This function is very similar to that derived for the SAR cross-section model. Its estimate follows a similar procedure. As the estimators of  $\beta$  and  $\sigma^2$  are a function of  $\rho$ , Elhorst 2003

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

$$\text{Log}L_C = C + T \log|I_n - \rho W| - \frac{NT}{2} ((u_0^* - \rho u_1^*)' (u_0^* - \rho u_1^*)) \quad (7.23)$$

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

Modelling spatial autocorrelation through a spatially autocorrelated error term only modifies the estimate of  $\sigma^2$  (the estimate of  $\beta$  is not affected). The generalised least squares method makes it possible to identify an estimator of  $\sigma^2$  if  $\lambda$  was known. In general, this is not the case and the estimation needs to be carried out again iteratively  $\beta$ ,  $\lambda$  followed by  $\sigma^2$ . The concentrated likelihood function can be maximised using residues ( $\varepsilon_{it}^*$ ) of the regression of  $y_{it}^*$  on  $x_{it}^*$ :

$$\text{Log}L_C = T \log(I_N - \lambda W) - \frac{NT}{2} \log(\varepsilon_{it}^* (I_N - \lambda W)' \varepsilon_{it}^* (I_N - \lambda W)) \quad (7.24)$$

Lee et al. 2010b have challenged this approach by showing that it does not necessarily make it possible to find consistent estimators of coefficients and standard deviations. The size of the bias and the parameters affected differs depending on the case. For example, when the model contains an individual fixed effect,  $\sigma^2$  is biased for large  $N$  and fixed  $T$ . If the model includes both time and individual effects,  $\beta$  and  $\sigma^2$  will be biased for  $N$  and large  $T$ . Based on these results, Lee et al. 2010b suggest corrections specific to each case to obtain consistent estimators from the direct approach. These corrections are available in the main econometrics software tools. We refer readers to Lee et al. 2010b and Elhorst 2014b for further details on this approach.

#### **Box 7.2.2 — Estimating a fixed effects model using the generalised method of moments.**

An alternative estimation strategy is based on the generalised method of moments. In spatial models, the strategy proposed by Kelejian et al. 1999 for cross-sectional data is extended to panel data by Kapoor et al. 2007 and Murti et al. 2011.

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

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

#### **7.2.2 Random effects model**

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

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

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

It can be noted that if  $\phi = 0$ , then the transformation *within* applies, and the random-effects model amounts to a fixed-effect model.

In a model without spatial autocorrelation, the likelihood function is written:

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

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

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

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

$$\text{Log}L_C = -\frac{NT}{2} \log [(e^o(\phi))' (e^o(\phi))] + \frac{N}{2} \log(\phi^2) \quad (7.28)$$

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

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

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

$$\begin{aligned} \text{Log}L = & -\frac{NT}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log|V| + (T-1) \sum_{i=1}^N \log|B| \\ & - \frac{1}{2\sigma^2} e' (\bar{J}_T \otimes V^{-1}) e - \frac{1}{2\sigma^2} e' (E_T \otimes (B'B)) e \end{aligned} \quad (7.29)$$

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

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

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

#### **Box 7.2.4 — Estimating a random effects model using the generalised method of moments.**

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

## 7.3 Specification tests

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

### 7.3.1 Choosing between fixed and random effects

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

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

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

$$S_{Hausman} = NT(\hat{\beta}_{MCG} - \hat{\beta}_{within})'(\hat{\Sigma}_{within} - \hat{\Sigma}_{MCG})^{-1}(\hat{\beta}_{MCG} - \hat{\beta}_{within}) \quad (7.30)$$

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

### 7.3.2 Specification tests for spatial effects

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

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

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

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

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

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

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

## 7.4 Empirical application

### 7.4.1 The model

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

$$p_{it} = b_0 + b_1 q_{it} + \varepsilon_{it} \quad (7.31)$$

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

capital between spatial units. The extensive specification of Verdoorn resulting from these analyses is<sup>4</sup>:

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

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

This specification is defined with R as follows:

---

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

---

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

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

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

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

where:

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

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

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

---

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

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

---

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

---

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

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

---

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

---

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

---

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

---

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

### 7.4.2 Data and spatial weights matrix

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

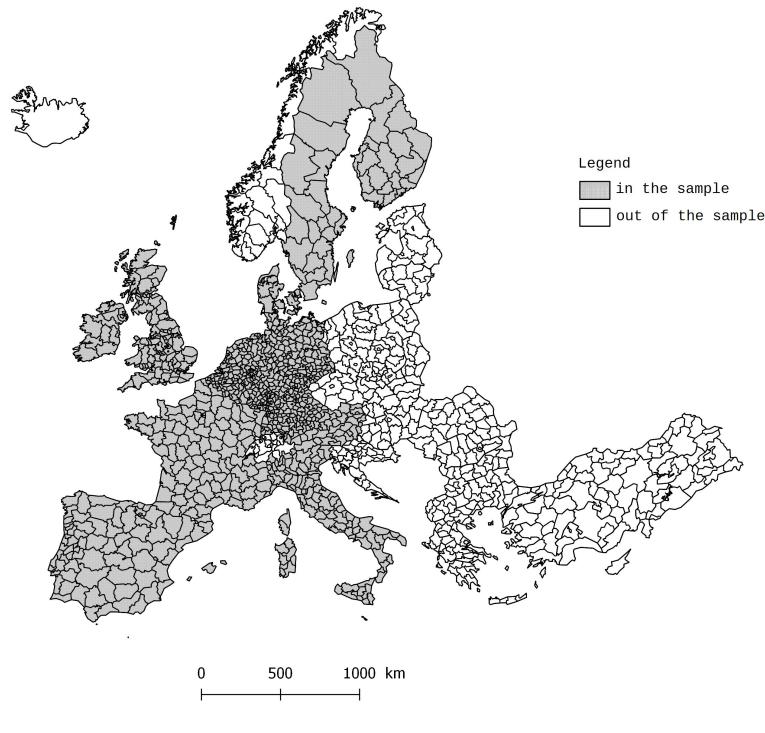


Figure 7.1 – Perimeter of the study

---

```
# Import data
data_panel <- read.csv("panel_average_3_years_1991_2008.csv", sep=";")

# Import shapefile (Gisco) as a "SpatialPolygonDataFrame"
shape_nuts3<-readShapeSpatial("NUTS_RG_60M_2006")

# Select NUTS3 (by NUTS3 level)
shape_nuts3<- shape_nuts3[shape_nuts3$STAT_LEVL_ == 3,]

# Select NUTS3 from the sample
data_panel_code<- data_panel[, "NUTS3"]
shape_nuts3<- shape_nuts3[shape_nuts3$NUTS_ID %in% data_panel_code,]

# Visualising the sample
plot(shape_nuts3)
```

---

In order to generate a table of descriptive statistics in L<sup>A</sup>T<sub>E</sub>Xformat of the explained variables and the explanatory variables of the model, it is possible to use package **stargazer** and apply the **stargazer** command on the database including the model variables. The result is shown in Table 7.2.

---

```
library(stargazer)
```

---

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

---

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

Table 7.2 – Descriptive statistics

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

---

```
# Creation of a k matrix plus close neighbours, k = 10
map_crd <- coordinates(shape_nuts3)
Points_nuts3 <- SpatialPoints(map_crd)
nuts3.knn_10 <- knearneigh(Points_nuts3, k=10)
K10_nb <- knn2nb(nuts3.knn_10)
wknn_10 <- nb2listw(K10_nb, style="W")
```

---

### 7.4.3 The results

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

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

---

```
# Table 7.3: estimation without consideration for spatial autocorrelation
summary(verdoorn_pooled <- plm(verdoorn, data = data_panel, model =
  "pooling"))
summary(verdoorn_fe1<- plm(verdoorn, data = data_panel,
  model = "within", effect="individual"))
```

---

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

---

Model:	p		
	pooled data	fixed effects ( <i>within</i> )	random effects (GLS)
	(1)	(2)	(3)
q	0.692*** (0.009)	0.604*** (0.010)	0.701*** (0.010)
u	0.0001*** (0.00001)	-0.0002 (0.0002)	0.0001*** (0.00001)
G	0.0001 (0.0001)	0.002*** (0.0001)	0.0003*** (0.0001)
d	-0.008*** (0.003)	-0.182*** (0.005)	-0.033*** (0.003)
Constant	0.146*** (0.012)		0.228*** (0.014)
Observations	5 160	5 160	5 160
R <sup>2</sup> adjusted	0.523	0.587	0.552

---

Table 7.3 – Estimates without consideration for spatial autocorrelation

**Note:** \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

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

---

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

# Hausman test robust to spatial autocorrelation (splm)
print(spat_hausman_ML_SEM<-sphtest(verdoorn,data=data_panel,
                                      listw =wknn_10, spatial.model = "error", method="ML
                                      "))
## Hausman test for spatial models
## data: x
## chisq = 1263.8, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

---

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

---

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

---

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

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

# Test 3
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="rlml",
         model="within")
## Locally robust LM test for spatial lag dependence sub spatial error
## data: formula (within transformation)
## LM = 0.0025551, df = 1, p-value = 0.9597
```

```

## alternative hypothesis: spatial lag dependence

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

```

---

Model:	<i>p</i>			
	pooled data	fixed effects (MV)		fixed effects (MMG)
	(1)	Baltagi error	KKP error	(4)
<i>q</i>	0.716*** (0.017)	0.650*** (0.008)	0.650*** (0.008)	0.836*** (0.009)
<i>u</i>	0.0001*** (0.00001)	0.0001 (0.0002)	0.0001 (0.0002)	0.0001 (0.0002)
<i>G</i>	-0.0004*** (0.0001)	0.001*** (0.0001)	0.001*** (0.0001)	0.0003*** (0.0001)
<i>d</i>	-1.70*** (0.003)	-0.163*** (0.0005)	-0.163*** (0.0005)	-0.164*** (0.005)
Constant	0.2*** (0.02)			
$\lambda$		0.566*** (0.02)	0.566*** (0.02)	0.513*** (0.02)
Observations	5 160	5 160	5 160	5 160

Table 7.4 – Estimations of the pooled data model and fixed effects model with spatial autocorrelation of errors

**Note:**\* $p < 0.1$  ; \*\* $p < 0.05$  ; \*\*\* $p < 0.01$

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

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

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

---

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

# Likelihood Maximum estimation
summary(verdoorn_SEM_pool <- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="pooling"))

# Fixed-effect SEM
summary(verdoorn_SEM_FE<- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="within", effect="individual", spatial.
error="b"))
summary(verdoorn_SEM_FE<- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="within", effect="individual", spatial.
error="kkp"))

# Generalised moments method estimation
summary(verdoorn_SEM_FE_GM <- spgm(verdoorn, data=data_panel,
listw = wknn_10, model="within", moments="fullweights",
spatial.error = TRUE))
```

---

## 7.5 Extensions

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

### 7.5.1 Dynamic spatial models

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

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

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

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

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

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

$$\begin{bmatrix} \frac{\partial qE(y)}{\partial x_{1k}} & \dots & \frac{\partial qE(y)}{\partial x_{nk}} \end{bmatrix}_t = (I_N - \rho W_N)^{-1}(\beta_k I_N + \theta_k W_N) \quad (7.38)$$

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

$$\begin{bmatrix} \frac{\partial qE(y)}{\partial x_{1k}} & \dots & \frac{\partial qE(y)}{\partial x_{nk}} \end{bmatrix}_t = [(1 - \tau)I_N - (\rho + \eta)W_N]^{-1}(\beta_k I_N + \theta_k W_N) \quad (7.39)$$

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

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

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

### 7.5.2 Multidimensional spatial models

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

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

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

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

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

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

$$u_{ijt} = \alpha_i + \mu_{ij} + v_{ijt}, \quad (7.42)$$

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

### 7.5.3 Panel models with common factors

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

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

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

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

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

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

## Conclusion

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

## References - Chapter 7

- Angeriz, Alvaro, John McCombie, and Mark Roberts (2008). « New estimates of returns to scale and spatial spillovers for EU Regional manufacturing, 1986—2002 ». *International Regional Science Review* 31.1, pp. 62–87.
- Anselin, Luc, Julie Le Gallo, and Hubert Jayet (2006). « Spatial panel econometrics ». *The econometrics of panel data, fundamentals and recent developments in theory and practice*. Ed. by Dordrecht Kluwer. 3rd ed. Vol. 4. The address of the publisher: Matyas L, Sevestre P, pp. 901–969.
- (2008). « Spatial panel econometrics ». *The econometrics of panel data*. Springer, pp. 625–660.
- Bai, Jushan and Peng Wang (2016). « Econometric analysis of large factor models ». *Annual Review of Economics* 8, pp. 53–80.
- Baltagi, Badi H, Peter Egger, and Michael Pfaffermayr (2013). « A Generalized Spatial Panel Data Model with Random Effects ». *Econometric Reviews* 32.5, pp. 650–685.
- Baltagi, Badi H and Long Liu (2008). « Testing for random effects and spatial lag dependence in panel data models ». *Statistics & Probability Letters* 78.18, pp. 3304–3306.
- Baltagi, Badi H, Heun Song Seuck, and Won Koh (2003). « Testing panel data regression models with spatial error correlation ». *Journal of econometrics* 117.1, pp. 123–150.
- Baltagi, Badi H, Seuck Heun Song, and Byoung Cheol Jung (2001). « The unbalanced nested error component regression model ». *Journal of Econometrics* 101.2, pp. 357–381.
- Baltagi, Badi H et al. (2007). « Testing for serial correlation, spatial autocorrelation and random effects using panel data ». *Journal of Econometrics* 140.1, pp. 5–51.
- Belotti, Federico, Gordon Hughes, Andrea Piano Mortari, et al. (2017b). « XSMLE: Stata module for spatial panel data models estimation ». *Statistical Software Components*.
- Bhattacharjee, Arnab and Sean Holly (2011). « Structural interactions in spatial panels ». *Empirical Economics* 40.1, pp. 69–94.
- Debarsy, Nicolas and Cem Ertur (2010). « Testing for spatial autocorrelation in a fixed effects panel data model ». *Regional Science and Urban Economics* 40.6, pp. 453–470.
- Debarsy, Nicolas, Cem Ertur, and James P LeSage (2012). « Interpreting dynamic space–time panel data models ». *Statistical Methodology* 9.1, pp. 158–171.
- Elhorst, J Paul (2003). « Specification and estimation of spatial panel data models ». *International regional science review* 26.3, pp. 244–268.
- (2012). « Dynamic spatial panels: models, methods, and inferences ». *Journal of geographical systems* 14.1, pp. 5–28.
- (2014a). « Matlab software for spatial panels ». *International Regional Science Review* 37.3, pp. 389–405.
- (2014b). « Spatial panel data models ». *Spatial Econometrics*. Springer, pp. 37–93.
- Ertur, Cem and Antonio Musolesi (2015). « Weak and Strong cross-sectional dependence: a panel data analysis of international technology diffusion ». *SEEDS Working Papers 1915*.
- Fingleton, Bernard (2000). « Spatial econometrics, economic geography, dynamics and equilibrium: a ‘third way’? ». *Environment and planning A* 32.8, pp. 1481–1498.
- (2001). « Equilibrium and economic growth: spatial econometric models and simulations ». *Journal of regional Science* 41.1, pp. 117–147.
- Fingleton, Bernard and John SL McCombie (1998). « Increasing returns and economic growth: some evidence for manufacturing from the European Union regions ». *Oxford Economic Papers* 50.1, pp. 89–105.
- Hausman, Jerry (1978). « Specification Tests in Econometrics ». *Econometrica* 46.6, pp. 1251–1271.
- Hsiao, Cheng (2014). *Analysis of panel data*. 54. Cambridge university press.

- Kapoor, Mudit, Harry H Kelejian, and Ingmar R Prucha (2007). « Panel data models with spatially correlated error components ». *Journal of Econometrics* 140.1, pp. 97–130.
- Kelejian, Harry H and Ingmar Prucha (1998). « A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances ». *Journal of Real Estate Finance and Economics* 17, pp. 99–121.
- (1999). « A generalized moments estimator for the autoregressive parameter in a spatial model ». *International Economic Review* 40.2, pp. 509–533.
- Krugman, Paul (1999). « The role of geography in development ». *International regional science review* 22.2, pp. 142–161.
- Le Gallo, Julie and Alain Pirotte (2017). « Models for Spatial Panels ».
- Lee, Lung-fei and Jihai Yu (2010a). « A spatial dynamic panel data model with both time and individual fixed effects ». *Econometric Theory* 26.2, pp. 564–597.
- (2010b). « Some recent developments in spatial panel data models ». *Regional Science and Urban Economics* 40.5, pp. 255–271.
- (2015). « Spatial panel data models ».
- Lesage, James and Robert K Pace (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.
- Manski, Charles F (1993a). « Identification of Endogenous Social Effects: The Reflection Problem ». *Review of Economic Studies* 60.3, pp. 531–542.
- Millo, Giovanni (2014). « Maximum likelihood estimation of spatially and serially correlated panels with random effects ». *Computational Statistics and Data Analysis* 71, pp. 914–933.
- Millo, Giovanni and Gianfranco Piras (2012). « splm: Spatial panel data models in R ». *Journal of Statistical Software* 47.1, pp. 1–38.
- Mutl, Jan and Michael Pfaffermayr (2011). « The Hausman test in a Cliff and Ord panel model ». *The Econometrics Journal* 14.1, pp. 48–76.
- Pesaran, M Hashem and Elisa Tosetti (2009). « Large panels with spatial correlations and common factors ». *Journal of Econometrics* 161.2, pp. 182–202.
- (2011). « Large panels with common factors and spatial correlation ». *Journal of Econometrics* 161.2, pp. 182–202.
- Piras, Gianfranco (2014). « Impact estimates for static spatial panel data models in R ». *Letters in Spatial and Resource Sciences* 7.3, pp. 213–223.
- Shi, Wei and Lung-fei Lee (2017a). « Spatial dynamic panel data models with interactive fixed effects ». *Journal of Econometrics* 197.2, pp. 323–347.
- (2017b). « A spatial panel data model with time varying endogenous weights matrices and common factors ». *Regional Science and Urban Economics*.
- Vega, Solmaria Halleck and J Paul Elhorst (2016). « A regional unemployment model simultaneously accounting for serial dynamics, spatial dependence and common factors ». *Regional Science and Urban Economics* 60, pp. 85–95.
- Verdoorn, JP (1949). « On the factors determining the growth of labor productivity ». *Italian economic papers* 2, pp. 59–68.
- Yu, Jihai, Robert De Jong, and Lung-fei Lee (2008). « Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large ». *Journal of Econometrics* 146.1, pp. 118–134.



# 8. Spatial smoothing

LAURE GENESES, AURIANE RENAUD ET FRANÇOIS SÉMÉCURBE  
INSEE

---

<b>8.1</b>	<b>Kernel smoothing</b>	<b>206</b>
8.1.1	Origin and formalism of kernel smoothing . . . . .	206
8.1.2	Adjusting to edge effects . . . . .	210
8.1.3	Choosing bandwidth . . . . .	211
<b>8.2</b>	<b>Geographical smoothing</b>	<b>212</b>
8.2.1	Smoothing weighted data . . . . .	212
8.2.2	Application using non-parametric regression . . . . .	215
8.2.3	Application using a non-parametric conditional density estimate .	216
8.2.4	Application using quantile smoothing . . . . .	217
<b>8.3</b>	<b>Implementation with R</b>	<b>217</b>
8.3.1	Under R, with package <i>spatstat</i> . . . . .	219
8.3.2	Under R, with package <i>btb</i> . . . . .	220
8.3.3	Optimal bandwidth tests . . . . .	223

---

## Abstract

Kernel smoothing is one of the key methods for analysing data and spatial organisation. The idea consist in filtering information to reveal underlying spatial structures.

From a conceptual point of view, kernel smoothing is a non-parametric estimation method of the intensity function of a point process with values in  $\mathbb{R}^2$ , based solely on one of its realisations (which has been observed). The theoretical intensity function in one point  $x$  is found by calculating the average points observed per unit surface on neighbourhoods containing  $x$ , these neighbourhoods being increasingly smaller.

However, in practice, there is only one (observed) realisation, and this approach, consisting of changing to the limit no longer makes sense. The non-parametric kernel methods circumvent this limitation, not by directly suggesting an estimation of the intensity function but by suggesting a smoothed estimation of it. Notwithstanding this approximation, when the bandwidth parameter is well chosen, the resulting estimates are statistically robust and geographically relevant, and make it possible to detect whether the intensity function is constant or variable in space.

Spatial analysis tools are used to produce appropriate geographical analyses. The aim is to develop simplified, clear mapping, relieved of the arbitrariness of territorial boundary lines, as well as partly mitigating the "Modifiable Area Units Problem". In this case, the bandwidth is a geographic generalisation parameter that maintains or deletes, depending on the requirements of the analysis and the details of the geographical phenomena observed. In practice, it is possible to smooth weighted data according to Brunsdon et al. 2002 — each point in the space is assigned a numerical value. Multiple types of smoothing can be carried out, including "classic" smoothing, based on local calculations of averages, or "quantile" smoothing using local calculations of quantiles

(median, decile), see Brunsdon et al. 2002. In addition, operations on smoothed values make it possible, in particular, to calculate “smoothed” ratios, such as the percentage of a sub-population within the population as a whole.

It has become fairly easy to implement smoothing, in particular using R software, for which several packages include functions that make such smoothing possible.

## 8.1 Kernel smoothing

The theoretical intensity function at point  $x$  is found by calculating the average of the points observed per surface on neighbourhoods containing  $x$  (see chapter 4: “Point configuration”) smaller and smaller point configurations. Kernel smoothing is a non-parametric estimation method for the intensity function of a point process with values in  $\mathbb{R}^2$  based solely on one of its realisations. To find the theoretical intensity function based on a single known realisation, it is not the intensity function itself that is estimated, but a function thereof.

From a practical point of view, kernel smoothing is a local modelling based on a selection of parameters.

**The kernel** describes how the neighbourhood is approached.

**The bandwidth** is the fundamental parameter in the analysis. It quantifies the «size» of the neighbourhood. This parameter results from a bias-variance trade-off between the spatial accuracy of the analysis and its statistical quality.

**The way in which edge effects** are handled explains how the geographical boundaries and the limits of observation territory are taken into account in the analysis.

Furthermore, a set of geographic coordinates can be set out, for which the smoothed values will be estimated (possibly different from all the geographical coordinates of the original data). Most of the applications made by INSEE smooth the data on tile grids (the new coordinate being the centre of the tile).

In this chapter, we will start out by discussing the foundations and formalism of kernel smoothing and then proceed to its implementation.

### 8.1.1 Origin and formalism of kernel smoothing

Historically, the first non-parametric intensity estimation method was based on building territorial intensity. It consists in calculating, for each territorial unit, the point intensity observed per surface unit. In this case, the intensity is also referred to as density. Within each of these territorial units, the estimated intensity is constant. For example, when calculating the density of a region, the said region is considered to be the same throughout the territory.

The practical interest of understanding intensity is based on the possibility of representing territorial densities in the form of choropleth maps, the first versions of which date back to the work of Baron Pierre Charles Dupin, see Palsky 1991. Geographers and statisticians then used this method to represent the distribution of the population within administrative regions. From a technical point of view, density maps generalise the histograms of the monodimensional analyses to the two-dimensional geographical areas. A sample density map is shown on Figure 8.1.

In the 20th century, geographers and statisticians gradually came to question the statistical and geographical relevance of this type of approach. Openshaw theorised its limits under the name *Modifiable Area Units Problem* (MAUP) . MAUP (see Figure 8.2) is broken down into two interdependent sub-problems — the scale effect and the zoning effect. The scale effect describes the dependence of the phenomenon observed on the average size of the spatial units. The larger the size, the smaller the local specificities and the more the analyses show the global structures. In

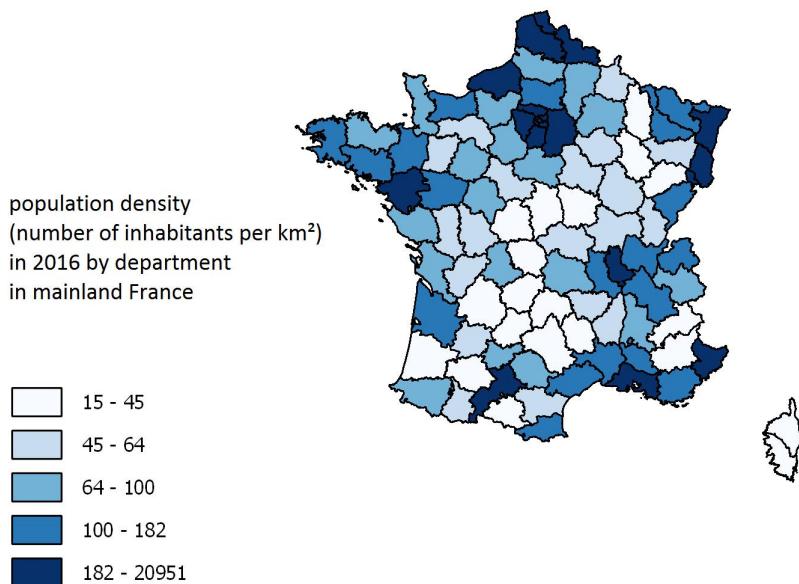


Figure 8.1 – Example of map density

Source: INSEE

contrast, small sizes retain local specificities and detail them. However, it should also be noted that the resulting analyses are sensitive to statistical noise, data quality and data accuracy. The zoning effect explains that the phenomena observed are dependant on the form of the spatial units. The concept of form includes the morphology of spatial units but also their position in space. Thus, if the spatial units' contours are displaced uniformly, the phenomenon observed is likely to be profoundly changed.

Kernel smoothing has inherited these reflections and aims to overcome the arbitrariness of territorial divisions. Kernel smoothing is rigorously defined in the context of spatial analysis, but similar methods in geography and statistics can be detected from the end of 19th century with the work of Louis-Léger Gauthier and Victor Turquan. This proximity (or entanglement) between the approach of spatial statisticians and the approach of geographers, justifies this chapter's focus on smoothing both from the standpoints of pure spatial analysis and a more operational geographical analysis.

In practice, the challenge for statisticians lies in observing only a single realisation. In concrete terms, to circumvent the challenge of estimating an intensity function based on a single realisation, kernel smoothing does not directly estimate that realisation, but a smoothed version obtained by convolution with a kernel  $K_h$ :

$$(K_h * \lambda)(x) = \int_{\mathbb{R}^2} \lambda(t) K(x-t) dt \quad (8.1)$$

$$\text{with } K_h(u) = \frac{1}{h^2} K\left(\frac{u}{h}\right)$$

and  $K$  a symmetrical function  $\mathbb{R}^2$  in  $\mathbb{R}$  positive and of integral 1

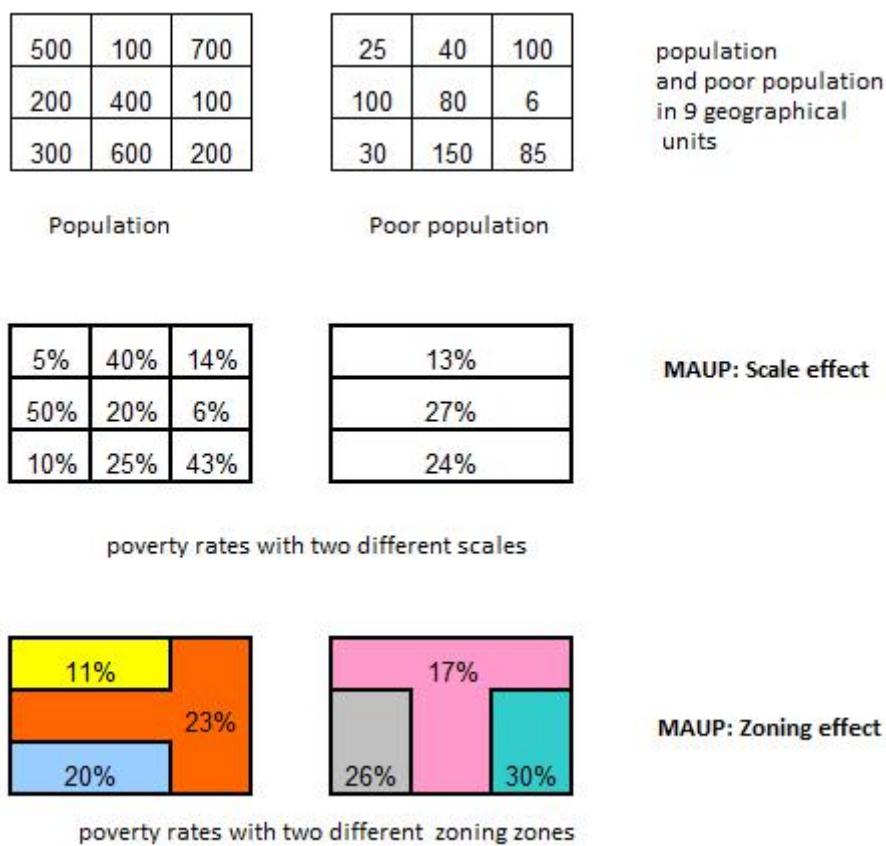


Figure 8.2 – MAUP Diagram: scale effect and zoning effect

**R** A simple metaphor for understanding the convolution operation consists in imagining that  $\lambda$  represents the distribution of the density of rabbit holes in space. Each hole is associated with a single rabbit. Each rabbit, to satisfy its needs, moves within a given proximity of its hole, so that its probability of being in position  $t$  in relation to its hole is  $K_h(t)$ . Convolution  $(K_h * \lambda)(x)$  in this case represents the local density of rabbits in  $x$ . If  $h$  is small, rabbits concentrate around their holes and the rabbit intensity function differs little from that of the holes. In contrast, if  $h$  is high, rabbits tend to mix in space and the rabbit intensity function is «blurred» compared to that of the holes.

To find an estimator for  $(K_h * \lambda)(x)$  from a set of points  $\{x_i\}$  resulting from the realisation of a point process, a simple idea consists in substituting the integral for  $\mathbb{R}^2$  by a sum on the points observed in the Equation (8.1).

**Definition 8.1.1 — Kernel smoothing.** Given  $K_h$  on a bandwidth kernel  $h$  and  $x$  a point  $\mathbb{R}^2$ , the estimated smoothed intensity in  $x$  is defined by:

$$\hat{\lambda}_h(x) = \sum_i K_h(x - x_i) \quad (8.2)$$

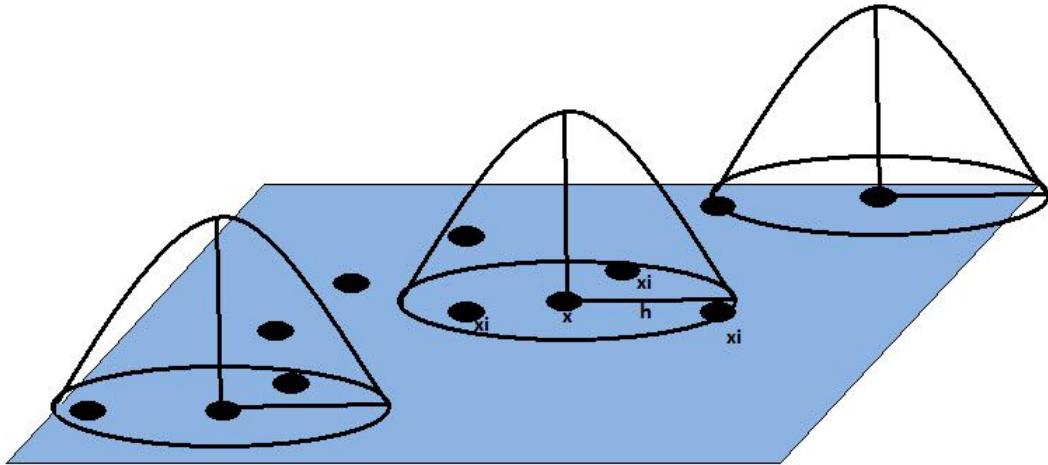


Figure 8.3 – Kernel smoothing scheme

**Note:** At each point of estimation lies a kernel function. The value of this function is highest at point level and decreases as one moves away.

$K_h$  in this formula plays a role similar to a territorial unit centred on each point in space  $\mathbb{R}^2$  with size  $h$ . In contrast to analyses based on a geographical split, the estimator in smoothed intensity controls the zoning effect of the MAUP, with the choice of kernel having little impact on the smoothing results. In contrast, the arbitrariness of the scale effect is maintained through the choice of bandwidth. Different kernels have been proposed in the literature. The most frequently used kernels are listed below.

**Definition 8.1.2 — Common kernels.**  $x$  is a point of  $\mathbb{R}^2$ .  $K^N$  and  $K^B$  are respectively referred to as the Gaussian kernel and quadratic kernel:

$$K_h^N(x) = \frac{1}{2\pi} e^{-\left\|\frac{x}{h}\right\|^2} \quad (8.3)$$

$$K_h^B(x) = \frac{9}{16} \cdot 1_{\|x\| < h} \cdot (1 - \left\|\frac{x}{h}\right\|^2)^2 \quad (8.4)$$

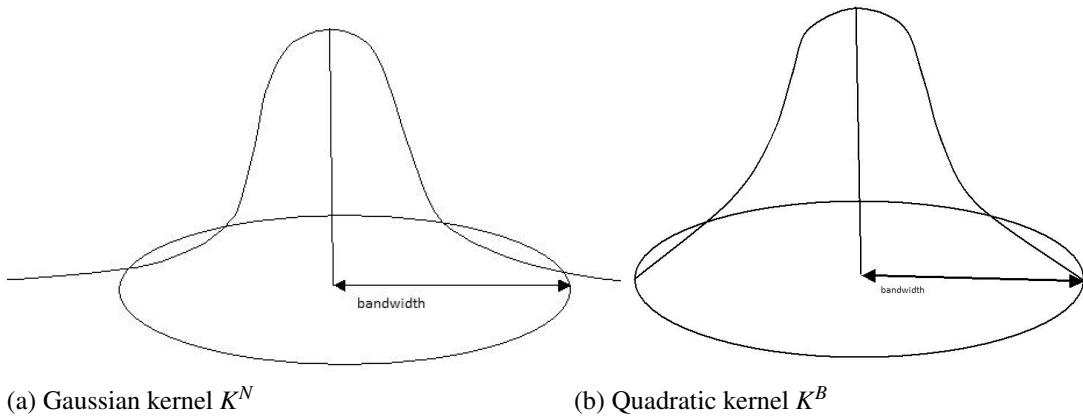


Figure 8.4 – Gaussian and quadratic kernels

**Note:** The quadratic kernel  $K^B$  gives greater weight to the closest points than to the remote points. It cancels itself out beyond the smoothening radius. On the contrary, Gaussian kernel  $K^N$  takes all points in the study area into account.

### 8.1.2 Adjusting to edge effects

As regards kernel estimation of the probability densities, kernel smoothing methods are impacted by an additional problem that arises when edge effects are taken into account. In the case of density estimation, the estimate is made on  $\mathbb{R}^n$ . Where kernel smoothing is concerned, the observed points are generally contained in an analysis window  $W$ , or, concretely, a polygon.

The nature of the window's borders can be of two kinds. Firstly, the window may result from the information collection protocol. For example, during an archaeological excavation, only a restricted area is dug for reasons of costs and opportunities. In this case, the borders are not inherent to the process observed, and without further information it is reasonable to assume continuity of intensity inside and outside the window. Secondly, the window may be induced by geographic configurations that have an impact on the underlying process generating the set of points observed. In geography, rivers, reliefs and coasts are all borders that restrict the settlement of human activities. Beyond such borders, the intensity of the phenomenon observed is null.

The formula (8.2) is the estimation formula, without adjusting for edge effects. In end-effect, the analysis window is ignored.

The purpose of adjusting to edge effects is to take into account the impact of the border when estimating the intensity. A variety of solutions has been suggested to this end. They differ in their approach of the outside of the observation area and in how fast they are carried out (Baddeley, see Baddeley et al. 2015a).

**Definition 8.1.3 — Adjusting for edge effects.**  $x$  is a point of  $\mathbb{R}^2$ , whereby the uniform and Diggle estimates (see Diggle 2013) are found using the following formulas:

$$\text{uniform correction: } \widehat{\lambda}_h^U(x) = \frac{1}{e_{h(x)}} \sum_i K_h(x - x_i) \quad (8.5)$$

$$\text{Diggle's correction: } \widehat{\lambda}_h^D(x) = \sum_i \frac{1}{e_{h(x_i)}} K_h(x - x_i) \quad (8.6)$$

where  $e_h(u) = \int_W K_h(u - v) dv$

When the analysis window is independent of the underlying process, the uniform estimate ensures continuity of intensity between the inside and outside of the window. However, if the intensity outside the window is deemed to be null, it is more opportune to use Diggle's estimation method, see Diggle 2013, which is conservative. In this case, the intensity integral estimated in the analysis window exactly matches the number of points observed. From an algorithmic point of view, Diggle's estimation method requires significantly more calculation time than the uniform estimate.

**R** The term  $e_{h(x)}$  can be interpreted as the intersection probability between two sets. Let us assume, again using our rabbit metaphor, that the window's spatial footprint matches that of an enclosure.  $K_h(x - u)$  approximately describes a rabbit's exploration territory around a hole found in  $x$  if the rabbit does not encounter any obstacles.  $e_h(x) = \int_W K_h(u - x) du$  is the part of the territory explored by the rabbit contained in the spatial footprint of the enclosure.  $e_{h(x)}$  is strictly lower than 1 if  $x$  immediately neighbours the border of the enclosure. However, if the "natural" exploration zone of the rabbits occupying the whole is entirely contained in the enclosure,  $e_{h(x)}$  is equal to 1.

In formula (8.5) the term  $e_{h(x)}$  is applied overall to the density estimate. Near the boundary of the enclosure, this term makes it possible to restore the estimated intensity. The closer the point is to the border, the lower  $e_{h(x)}$  and the greater the compensation will be. Uniform correction considers that at the window's boundary, the distribution of rabbit holes is almost homogeneous inside and outside the window. Intuitively, this correction amounts to postulating that the enclosure has no effect on the mobility of rabbits, which walk across its boundary without realising. More specifically, it is considered that rabbits whose holes are located outside the enclosure also contribute to the intensity calculated within the enclosure.

Diggle's estimation method, formula (8.6), assumes that the window is an integral part of the properties of the underlying process. In other words, the enclosure forms an insurmountable boundary for the rabbits and all the rabbits are contained in the enclosure. By dividing  $K_h(x - x_i)$  with the term  $e_{h(x_i)}$ , we ensure that the rabbit of hole  $i$  has a probability equal to 1 of remaining in the spatial footprint of the enclosure.

### 8.1.3 Choosing bandwidth

The choice of bandwidth determines the extent to which the estimation of intensity function will be "smoothed". In spatial analysis, bandwidth results from a bias-variance compromise. The bias is caused by the fact that the intensity function estimator does not directly estimate the intensity function but a smoothed version thereof. The greater the bandwidth, the greater the bias. The variance, on the contrary, decreases according to the bandwidth. The greater the bandwidth, the greater the number of points involved in calculating local estimations, which tends to reduce the estimation variance.

Several methods are available to automatically suggest a bandwidth that minimizes an error criterion. As the desired intensity function is obviously not available, some of these methods are based on cross-validation methods. They use the point distribution observed, and assume that it follows a Poisson distribution in order to estimate optimal bandwidth. In section 8.3, examples using the package's cross-validation functions *spatstat* of R will be suggested. These examples highlight the high variability of the proposed bandwidths based on the selected error criteria. Furthermore, the existence of a single bandwidth relevant for the entire extent of the zone studied is a central hypothesis. Several adaptive smoothing methods have been suggested to overcome this limit. Readers will find interesting material in Baddeley's book (Baddeley et al. 2015a) on this topic, which makes use of package *spatstat* in R.

In fact, no bandwidth is optimal: all are capable of providing an apposite depiction of the world as defined in the MAUP. Some geographers recommend adopting a multi-scale approach to study the multiplicity of spatial aspects within a single phenomenon.

## 8.2 Geographical smoothing

Geographical smoothing is based on the intensity estimation method presented above. It is not intended to calculate intensities, but to come up with simplified mapping representations. The principle of this form of use in geography is to represent not the value observed at a single point, but a weighted average of the values observed in the neighbourhood of this point, within a predefined radius.



Smoothing can be interpreted as a tool capable of guaranteeing a form of **confidentiality**. It makes it possible to represent initially *ad hoc* and confidential data in aggregate form. It is nevertheless important to remain vigilant regarding the number of points used to produce the smoothed estimate.

### 8.2.1 Smoothing weighted data

In this case, each point  $x_i$  is assigned a numerical value  $w_i$ . For example,  $x_i$  can represent a housing unit, and  $w_i$  the number of inhabitants in this housing unit. For this, we need only (see Brunsdon et al. 2002) use a weighted version of the kernel estimators described above. In formula (8.2), weight  $w_i$  is multiplied by the contribution of a point to the intensity estimator.

**Definition 8.2.1 — Weighted kernel estimators.** Given  $K_h$  a bandwidth kernel  $h$  and  $x_i$  one point on  $\mathbb{R}^2$  with assigned weighting  $w_i$ , the smoothed intensity estimated in  $x$  is defined by:

$$\hat{\lambda}_h(x) = \sum_i w_i K_h(x - x_i) \quad (8.7)$$

While the choice of the kernel  $K_h$  has little influence on the smoothing results (see Figure 8.5), the choice of bandwidth  $h$  is of fundamental importance, although fairly arbitrary.

As has been noted above, the bandwidth acts like a smoothing parameter, controlling the balance between bias and variance. A high radius leads to a significantly smoothed density and high bias. A small radius generates a density subject to little smoothing, with high variance. It is generally up to the user whether a compromise should be made, depending on the desired level of aggregation. It is advisable to test several bandwidth values, making it possible to reveal local variations at different scales. The maps in Figure 8.6 are examples of smoothed maps for Paris and its suburbs, with three different smoothing radii.

This estimate is valuable in that it focuses attention not on the points and their distribution, but on their environment. The bandwidth thus makes it possible to define this environment.



Multiple algorithms exist to determine a so-called "optimal" smoothing radius. These tests can yield various and sometimes very widely-differing results (see implementation). It is advised that they be used only for indicative purposes, and that the user choose the smoothing radius based on experience with the data and the issue.

Operations can be carried out on smoothed variables, in particular ratios. The theoretical rationale for this can be found on pp. 34 to 37 of Floch's working document (Floch 2012b). In practical terms, to find the smoothed value of the ratio of two variables, it is essential to separately calculate the smoothed values of the numerator and denominator, then to calculate the ratio between the smoothed value of the numerator and the smoothed value of the denominator. Do not directly calculate a smoothed ratio value. The map would be distorted, as the same importance would wrongly be given to all territories, despite their being unequally populated.

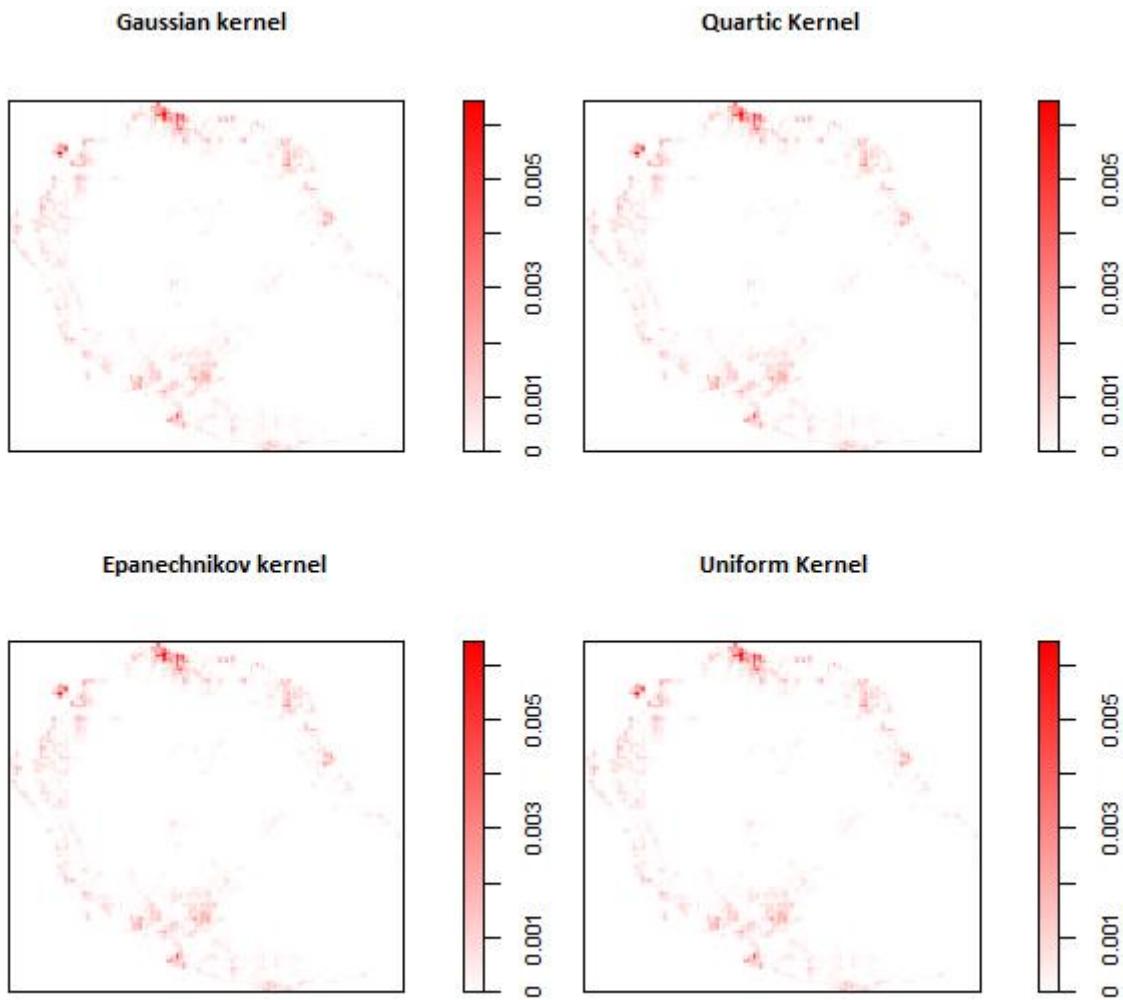


Figure 8.5 – Comparison of results found using four different kernels from the function `density.ppp` of package *spatstat*

**Source:** INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011

**Note:** The variable shown is the smoothed number of households on Reunion Island.

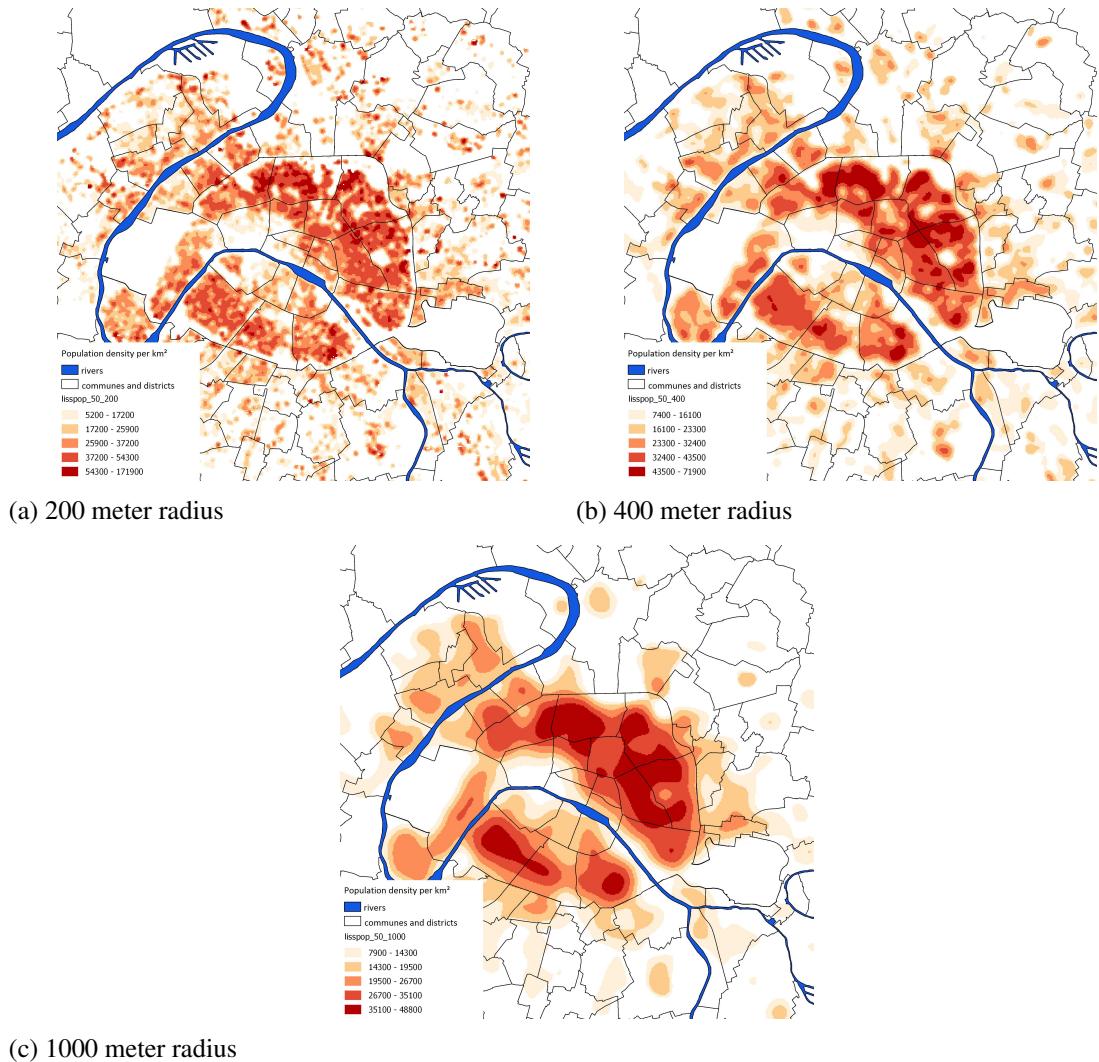


Figure 8.6 – Three different smoothing radii for population density in Paris and its suburbs — 200 meters, 400 meters, 1000 meters

**Source:** INSEE-DGFIP-CNAF-CNAV-CCMSA, *Localised social and tax file 2012*

**Note:** the tiles depicted contain more than 11 households.

### 8.2.2 Application using non-parametric regression

In this example, the focus is on **average income calculation** per person. We have two variables — income, and number of people. Average income is equal to the sum of the total income, divided by the sum of the number of people. Income and the number of people are smoothed separately. The ratio can then be calculated.

The maps are derived from Figure 8.7 as regards Paris and the surrounding municipalities in the immediately-surrounding suburbs (*i.e.* the three departments bordering Paris).

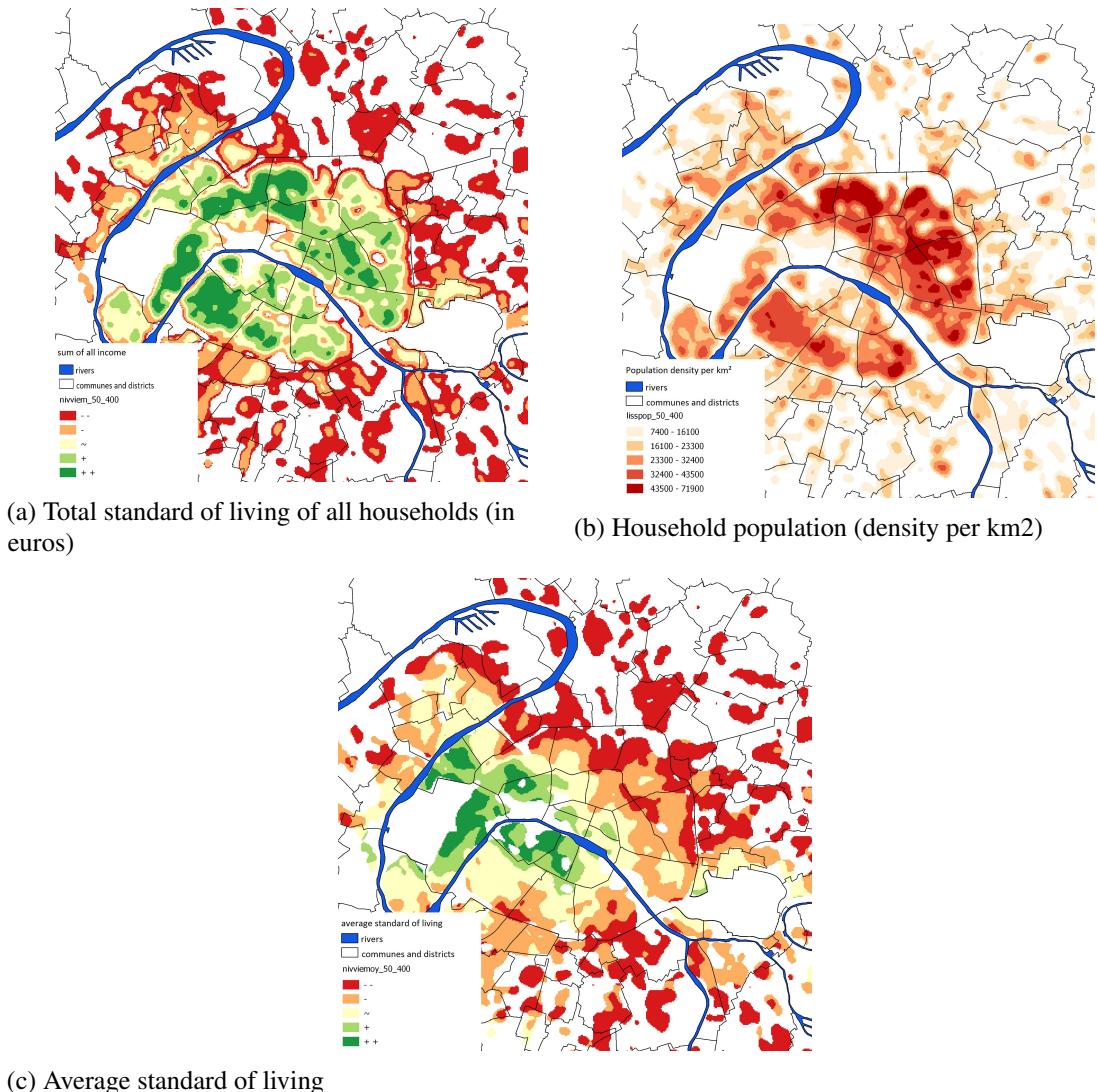


Figure 8.7 – Calculation of a smoothed average standard of living

**Source:** INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised social and tax file 2012

**Note:** The tiles shown contain more than 11 households. As to maps depicting levels of income, the markers “++”, “+” and “~”, “-” and “-” respectively reflect very high, high, average, low or very low values for the indicator in question. They were used for questions of non- profiling of the population.

Map 8.7a of the total standard of living of households does not give much information on its own. The total standard of living per square needs to be compared to population in each tile. On map 8.7a showing the number of people, the population is very dense within the municipality

of Paris, mainly north-east of the Seine, and to a lesser extent deeply southward.

On map 8.7a, the average standard of living per person is very high in the heart of Paris, essentially in the west.

- R** This is no longer a theoretical framework. This calculation is based on tools comparable to a non-parametric regression. Roughly speaking, it is as if a weighted geographical regression were carried out, limited to a single variable — the constant (see chapter 9: “Geographically Weighted Regression”).

### 8.2.3 Application using a non-parametric conditional density estimate

The focus is on the **proportion of poor households** across all households. The smoothed value of the number of poor households in a territory is calculated, and the smoothed value of the total number of households in the territory is calculated. The ratio is then calculated.

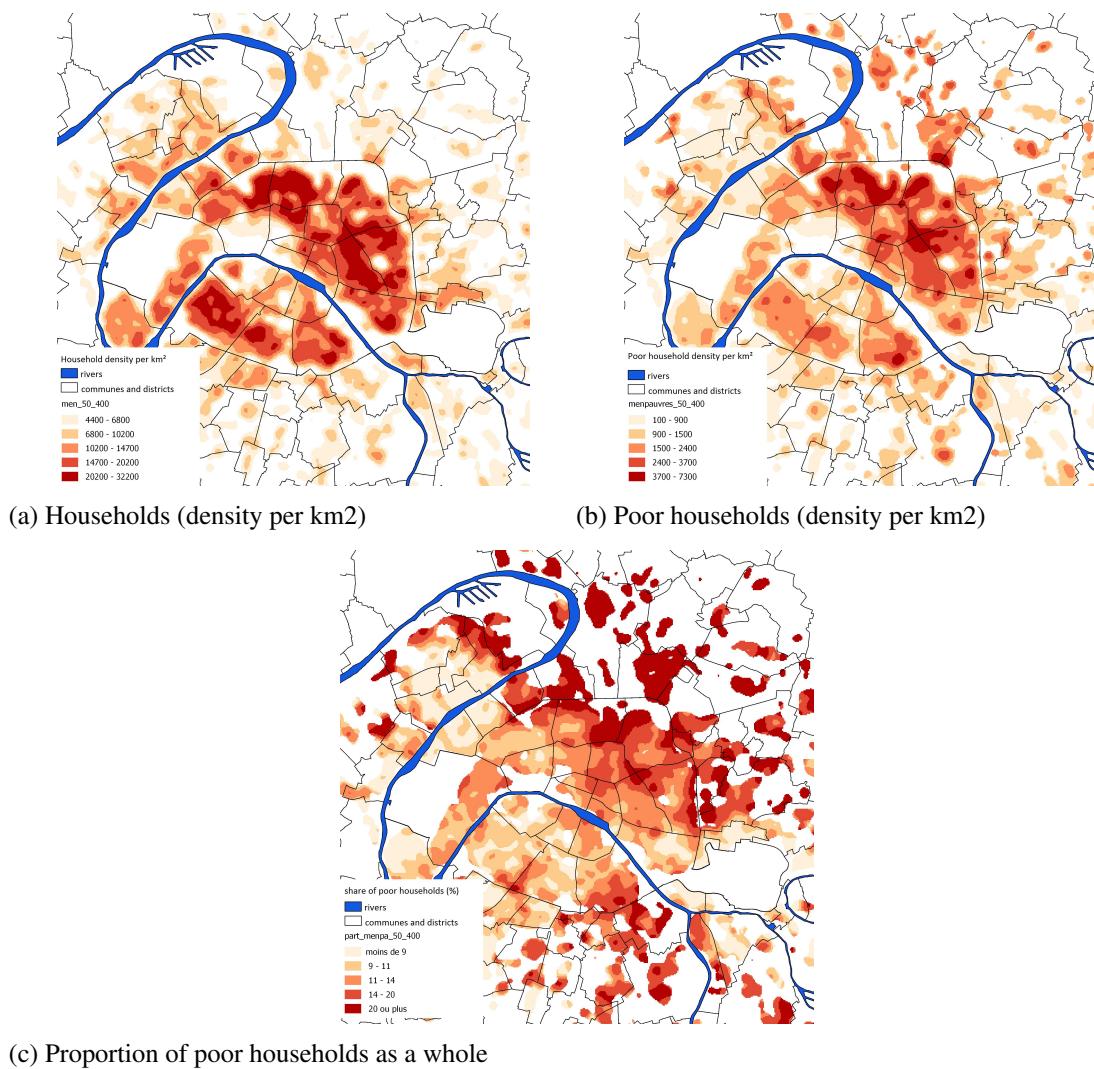


Figure 8.8 – Calculating the smoothed proportion of poor households

**Source:** INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised social and tax file 2012

**Note:** The tiles depicted contain more than 11 households.

According to the map from Figure 8.8a, the most populated areas are found in the heart of Paris, both in the north-east quarter and the south-west quarter. In Figure 8.8b, there are many poor

households in Paris, that tend to be located in the north-east quarter. In Figure 8.8c, the proportion of poor households in all households provides additional information. The map highlights the less densely populated areas, where the share of households living below the poverty line is nonetheless high. These are municipalities located north of Paris.

Thus, depending on the map produced, the messages derived can be different. When analysing the rates, it is also of fundamental importance to analyse the distribution of the number of people alone (population density, for example), in order to verify the robustness of the calculated rates and their representativeness.

- R This calculation can be likened to a conditional probability calculation. The result is a map of poverty rates at the local level, which is close to the idea of identifying the likelihood of a household's being poor, on the basis of its having settled in a given place.
- R **Important!** In theory, it is always possible to calculate the ratio of two smoothed variables. In practice, it is important to pay attention to the small numbers. In the example where the proportion of poor people was calculated, areas with small numbers could wrongly be shown separately in the smoothed map. Having a low population, they would not appear on a map showing raw data. Thus, by failing to take this phenomenon into account, we could mechanically give the - skewed - impression that all territories are populated.

#### 8.2.4 Application using quantile smoothing

The smoothing described up to this point is an average smoothing, in the sense that it is based on local calculations of averages. In the article by Brunsdon et al. 2002, the authors extend this concept, in order to define local statistics based on quantiles (median, deciles, etc.). These indicators are deemed, in the exploratory analysis of "traditional" data, to be less sensitive to extreme values. Quantile smoothing makes it possible above all to calculate indicators that considerably enrich the analysis of certain variables, in particular income variables.

The four thumbnails in Figure 8.9 represent multiple smoothed indicators calculated based on standard of living (source *INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised Social and Tax File 2012*), *i.e.* the disposable income of a household divided by the number of consumer units in the household.

The maps in Figure 8.9 are centred on Paris, and include the immediately surrounding suburbs. Smoothing is performed on 50-meter tiles, with a 400-meter bandwidth. Only the tiles for which the number of observations (of households) contributing to the estimate is strictly greater than 50 have been used for viewing.

Map 8.9a represents the median standard of living. Zones can be seen in the west where residents are much more affluent. Maps 8.9b and 8.9c show the 1st decile and 9th deciles of living standard. Map 8.9d depicts the interdecile ratio (ratio between the 9th and 1st deciles) and provides additional insight. Interdecile ratio is stated without units. It shows the minimum standard of living of the richest 10% relative to the maximum standard of living of the poorest 10% and brings out the gap between the top and bottom of the distribution. This is one of the measures of inequality in this distribution. In the aforementioned zones in the west, the interdecile ratios are very high. These neighbourhoods are home to populations with a very high standards of living, as well as populations with much lower standards of living.

### 8.3 Implementation with R

R offers multiple packages that can be used to perform smoothing. The practical implementation process is detailed below, using packages *spatstat* and *btb*, applied to data pertaining to Reunion

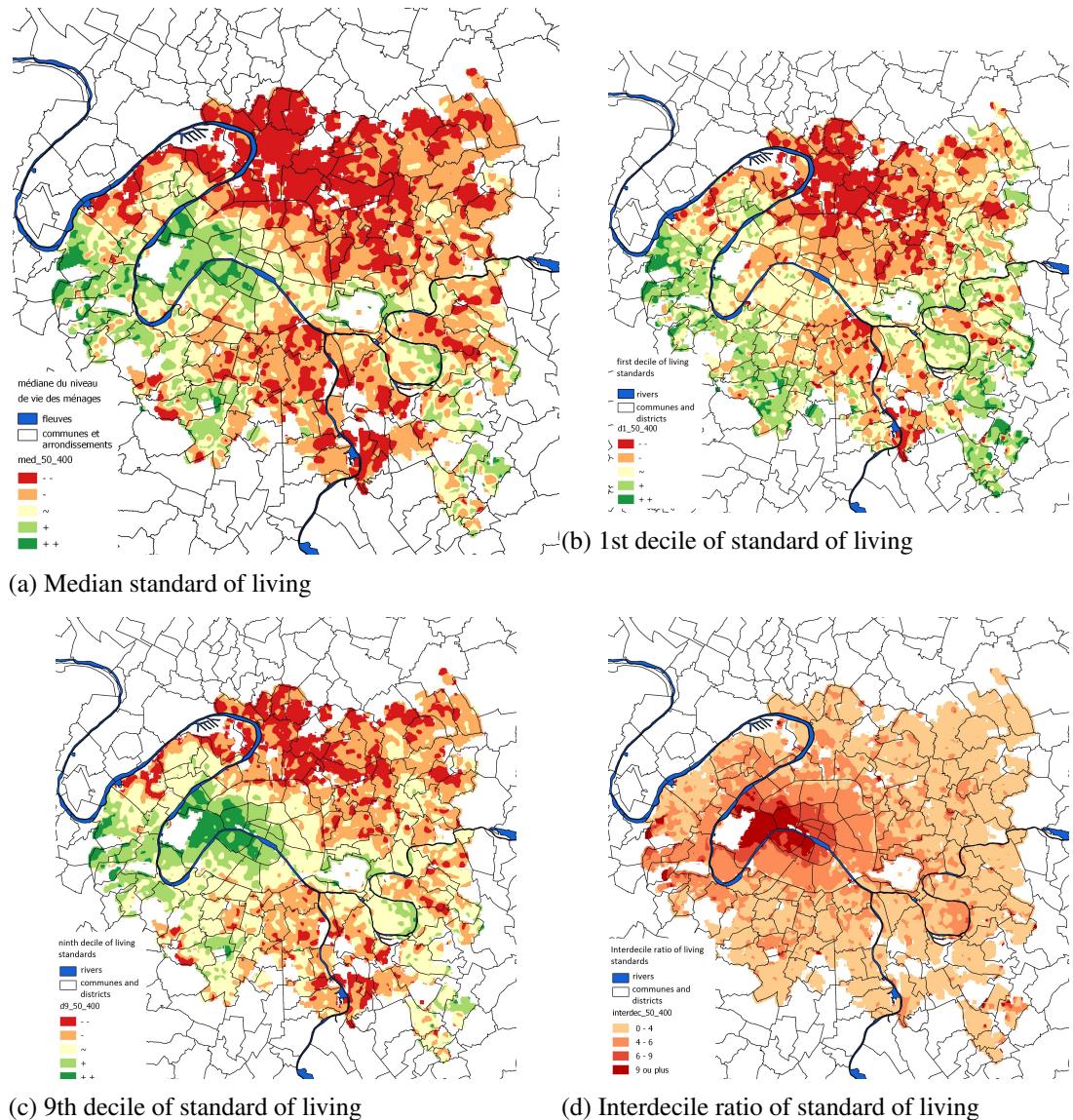


Figure 8.9 – Standard of living distribution

**Source:** INSEE-DGFIP-CNAF-CNAV-CCMSA, Localised social and tax file 2012

**Note:** The tiles shown contain more than 11 households. As to maps depicting levels of income, the markers “++”, “+” and “~”, “-” and “–” respectively reflect very high, high, average, low or very low values for the indicator in question. They were used for questions of non- profiling of the population.

Island. The data used in the example are the dataframe *reunion.Rdata* provided in package *btb*. An overview of this dataframe is provided in Figure 8.10.

	x	y	houhold	phouhold
1	359500	7634300	5.0693069	2.37623762
2	359500	7634500	26.9306931	12.62376238
3	355900	7634500	15.0000000	4.000000000
4	356100	7634500	39.0000000	20.000000000
5	356300	7634500	41.6428571	15.14285714
6	356500	7634500	2.3571429	0.85714286
7	359700	7634500	11.4210526	0.00000000
8	359700	7634700	2.5789474	0.00000000
9	359900	7634500	12.0000000	6.000000000
10	355700	7634700	1.0243902	0.00000000
11	355700	7635100	1.3658537	0.00000000
12	355700	7635300	11.6097561	0.00000000
13	355900	7634700	20.0000000	7.000000000
14	356100	7634700	131.0000000	71.000000000
15	356300	7634700	110.0000000	58.000000000

Figure 8.10 – The first 15 lines of the *data.frame* *reunion.Rdata* in package *btb*

**Source:** INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011

**Scope:** Reunion Island

This is 200-meter grid data, downloadable on *insee.fr*. The source is *INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011*.

The variables are thus defined:

- x: longitude (projection system: WGS 84 / UTM zone 40S, EPSG code: 32740)
- y: latitude (projection system: WGS 84 / UTM zone 40S, EPSG code: 32740)
- houhold: number of households
- phouhold: number of poor households (poverty definition at 60%)

### 8.3.1 Under R, with package *spatstat*

The R package known as *spatstat* is a comprehensive package dedicated to analysing spatial point processes. It is available on the CRAN website at the following address: <https://CRAN.R-project.org/package=spatstat>

The function *density.ppp* available in package *spatstat* makes it possible to run data smoothing. The use of this function requires the use of an object in format *.ppp* upon entry. To use this function, the x and y coordinates of the *data.frame* must be converted to *.ppp*.

---

#smoothing of the houhold variable (number of households) with Spatstat

```
library(spatstat)
```

---

```

library(btb)#only for the meeting dataframe
data(reunion)

# duplicate coordinate aggregation and deletion
base_temp <- aggregate(houhold ~ x+y, reunion, sum)

# x,y transformation into .ppp objects
base.ppp = spatstat::ppp(base_temp$x, base_temp$y,
c(min(base_temp$x), max(base_temp$x)),
c(min(base_temp$y), max(base_temp$y)) )

#density.ppp function call
#the sigma parameter is h/2 with h the bandwidth
densite <- spatstat::density.ppp (base.ppp, sigma = 200, weights=base_temp$houhold )

#map display
plot(densite, main = "Spatstat smoothing, user defined radius")

```

---

### 8.3.2 Under R, with package *btb*

Package *btb* ("beyond the border")<sup>1</sup> is online on the CRAN website, at the following address: <https://CRAN.R-project.org/package=btb>. It offers functions dedicated to urban analysis and implements a density estimate using the KDE method (kernel density estimator), *i.e.* a kernel method. The kernel used is a quadratic kernel.

In the estimation produced by the package, the edge effect is taken into account in the smoothing function *kernelSmoothing* via Diggle's correction (Diggle 2013). This correction makes it possible in particular to deal with the case of boundaries depicting geographical limits (coasts, for example). Within the observation area, the intensity is non-null. Outside the observation area, it is null. The method implemented is conservative (thanks to standardisation). Before and after smoothing, the number of points observed is the same.



Calculation times have been significantly reduced, in several ways:

- by coding in C++ all the most time-consuming methods;
- by limiting, for each point, to an observation window around this point, making it possible to limit the number of operations (calculations of distances) to be carried out.

---

```

#smoothing with btb: calculating the proportion of poor households
#the numerator (number of poor households), and the denominator (total
#number of households) are smoothed separately

```

```

library(btb)

#data loading
data(reunion)

#smoothing

```

---

1. There will soon be a version of package *btb* adapted to new package *sf* (simple features).

---

```

#parameter setting
pas <- 200 #200-meter square
rayon <- 400 #400-meter bandwith

#smoothing function call
#the function automatically smoothes all the variables contained in the
#database
#here, phouhold and houhold are smoothed

dfLisse <- btb::kernelSmoothing(dfObservations = reunion, iCellSize = pas,
                                  iBandwidth = rayon, sEPSG="32740")

#rate of poor households : ratio of smoothed variables
dfLisse$txmenpa = 100 * dfLisse$phouhold / dfLisse$houhold

#overview in R
library(sp)
library(cartography)
#map display
cartography::choroLayer(dfLisse, var = "txmenpa", nclass = 5, method =
  "fisher-jenks", border = NA, legend.pos = "topright", legend.title.txt =
  "txmenpa (%)")

#title and outline added
cartography::layoutLayer(title = "Reunion Island : rate of poor households"
  ,
  sources = "",
  author = "",
  scale = NULL,
  frame = TRUE,
  col = "black",
  coltitle = "white")

```

---

The resulting map is shown in Figure 8.11.

The user can also export the result in shapefile format, then rework it in a GIS.

---

```
#export in shapefile format

rgdal::writeOGR(as(dfLisse, 'Spatial'), "txmenpauvre.shp", "txmenpauvre",
  driver = "ESRI Shapefile")
```

---

Package *btb* also enables the use of quantile smoothing, described above. The user need only specify as a parameter *vQuantiles* the quantile vector to be calculated. For example *c(0.1, 0.25, 0.5)* will return the first decile, the first quartile and the median of each of the variables of the input *data.frame*.

---

```
# quantile smoothing
library(btb)
data(reunion)
```

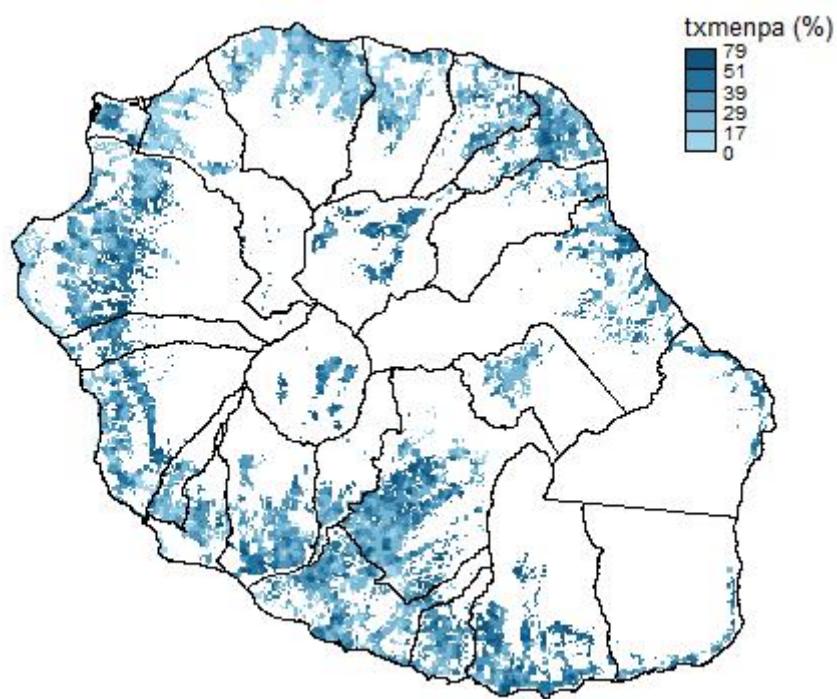


Figure 8.11 – Rate of poor households on Reunion Island after smoothing

**Source:** INSEE, Localised Tax Revenues as at 31/12/2010 and Housing Tax as at 01/01/2011

**Note:** Ratio between number of smoothed poor households and the total number of smoothed households. The black outlines show the municipal boundaries.

---

```
#parameter setting
pas <- 200
rayon <- 400

#smoothing function call
dfLisse_quantile<- btb::kernelSmoothing(dfObservations = reunion,
                                            iCellSize = pas,
                                            iBandwidth = rayon,
                                            vQuantiles = c(0.1, 0.5, 0.9),
                                            sEPSG="32740")

#export to QGIS
rgdal::writeOGR(as(dfLisse_quantile, 'Spatial'), "lissage_quantile.shp", "
  lissance_quantile",
  driver = "ESRI Shapefile")
```

---



Package *btb* defaults to an automatic tile grid. The smoothing function can also be queried using a grid of the user's choice. In this case, the user must have a `data.frame` consisting of two columns `x` and `y`, which match up with the desired centroid coordinates.

---

```
#smoothing function with grid, optional to user
kernelSmoothing(dfObservations, iCellSize, iBandwidth, dfCentroids)
```

---

### 8.3.3 Optimal bandwidth tests

In R, multiple methods proposing to calculate an "optimal" bandwidth can be implemented, based on different criteria. The aim is generally to minimise an error measurement. In *spatstat* for example, the following four functions are found: `bw.diggle`, `bw.ppl`, `bw.frac` and `bw.scott`.

#### With function `bw.diggle` in *spatstat*

Function `bw.diggle` in *spatstat* chooses a bandwidth that minimises a criterion  $M(\sigma)$  based on average quadratic error (MSE for *Mean Square Error*) in the estimator.

The chart in Figure 8.12 represents criterion  $M(\sigma)$ , which must be minimised. To find value  $\sigma$ , we will need to identify the value on the x-axis, which matches up with the minimum value on the y-axis.

For more details, see <https://www.rdocumentation.org/packages/spatstat/versions/1.49-0/topics/bw.diggle>.

---

```
#the base.ppp created above is used again
# bw.diggle test for optimal bandwidth
bw_diggle <- spatstat::bw.diggle(base.ppp)
plot(bw_diggle, main = "cross validation")

#density.ppp call with the automatically calculated bandwidth
densite_optim <- spatstat::density.ppp(base.ppp,bw_diggle, weights=base_
temp$houhold)
```

---

The result is:

---

```
bw_diggle
##     sigma
## 141.9445
```

---

Using the default settings, the proposed value for  $\sigma$  is 142 metres or 284 metres for bandwidth  $h$  ( $\sigma = h/2$ , see package documentation).

#### With function `bw.ppl` of **spatstat**

The bandwidth is chosen by calculating a maximum likelihood estimator, using a cross-validation method (*likelihood cross-validation criterion*). The calculations are then iterated. Each time, we work only on  $n - 1$  observations, then validate the model on the observation that had been discarded. We repeat this  $n$  times.

The graph below shows the CV criterion( $\sigma$ ) that we wish to minimise. To find value  $\sigma$ , we will need to identify the value on the x-axis, which matches up with the maximum value on the y-axis.

For more details, see <https://rdrr.io/cran/spatstat/man/bw.ppl.html>.

---

```
#the base.ppp created above is used again

# bw.ppl test for optimum bandwidth
bw_ppl <- spatstat::bw.ppl(base.ppp)
plot(bw_ppl, main = "bw.ppl")
```

---

The result is:

---

```
bw_ppl
##     sigma
## 286.0097
```

---

Using the default settings, the proposed value for the  $\sigma$  value is 286 metres.

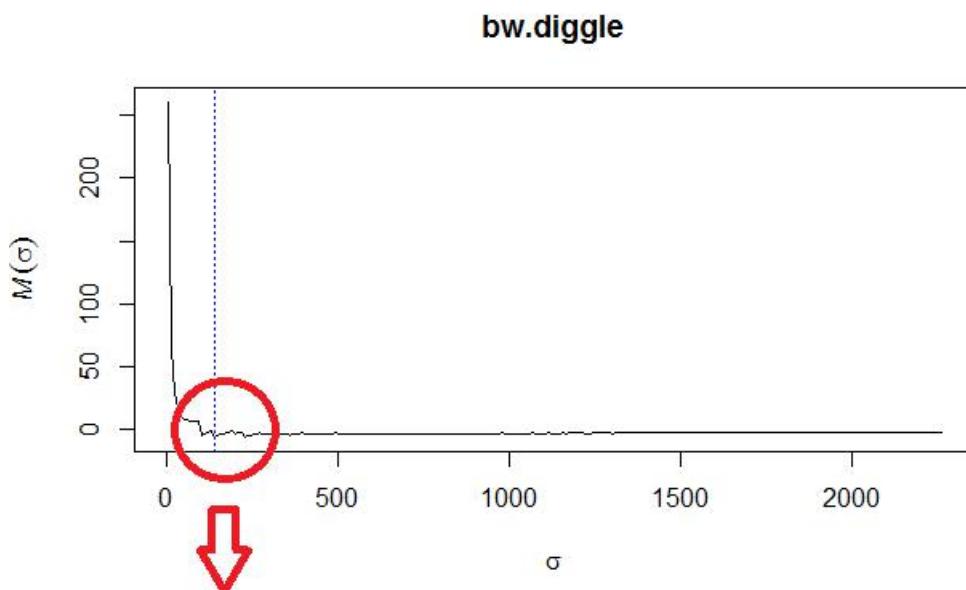
#### With function `bw.frac` of **spatstat**

This method selects a bandwidth based solely on the geometry of the observation window.

The bandwidth is a quantile (specified by the user) of the distance between two independent points, chosen randomly in the window. By default, the first distribution quartile is used. If  $CDF(r)$  is used to denote the cumulative distribution function of the distance between two independent points randomly and uniformly distributed in the window, then the value that is returned is the quantile with probability  $f$ . The bandwidth is then the  $r$  value, such that  $CDF(r)=f$ . First, the algorithm calculates the cumulative  $CDF(r)$  distribution function with the function `distcdf` of package *spatstat*. This function makes it possible to calculate the function  $CDF(r) = P(T \leq r)$  of the ECU  $T=|X_1-X_2|$  between two randomly-chosen independent points  $X_1$  and  $X_2$ . Then we look for the smallest number  $r$  such that  $CDF(r) \geq f$ .

The chart below shows the  $CDF(r)$  function. To obtain the bandwidth, we read the value of x-axis  $r$  such as  $CDF(r) = 0.25$  (by default, the first quartile is used).

For more details see <https://www.rdocumentation.org/packages/spatstat/versions/1.48-0/topics/bw.frac>.



By zooming in on the proposed value:

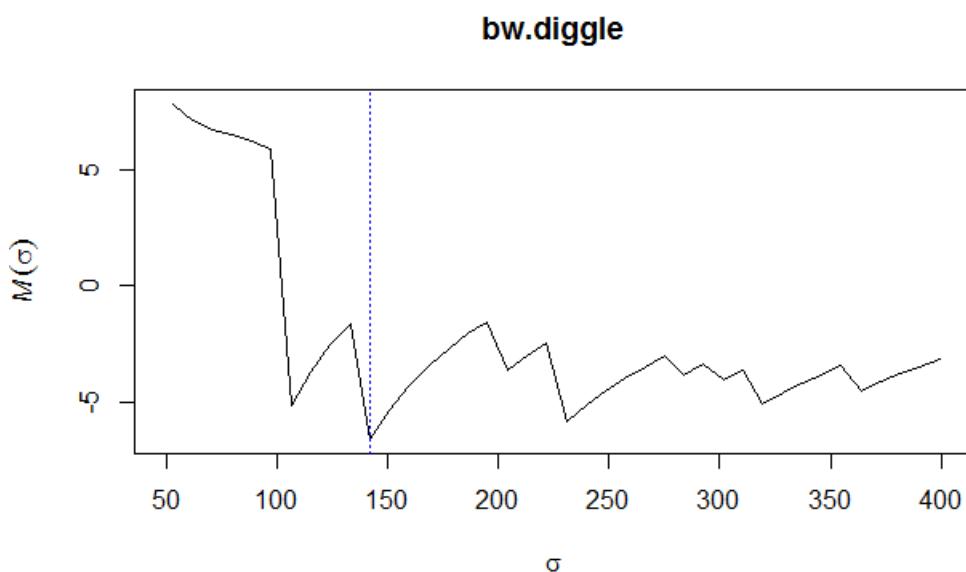
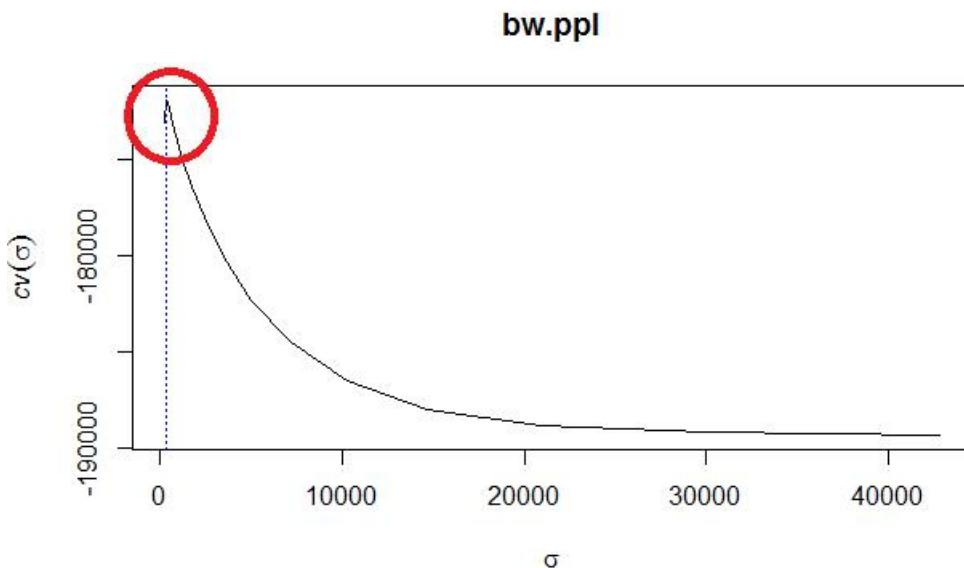


Figure 8.12 – Criteria  $M(\sigma)$  found using function `bw.diggle` of package *spatstat* in R

**Source:** INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011  
**Scope:** Reunion Island



By zooming in on the proposed value:

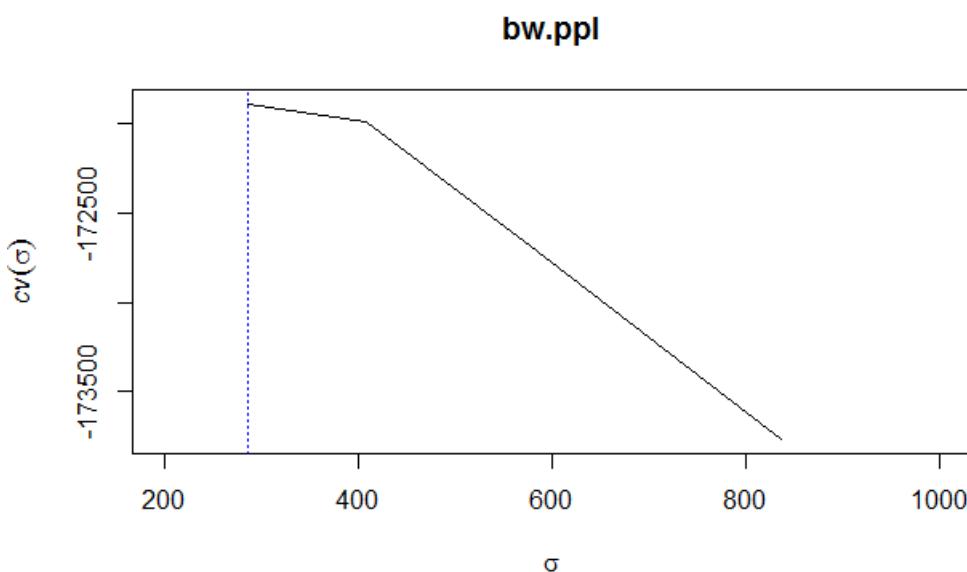


Figure 8.13 – The CV criterion( $\sigma$ ) derived using function `bw.ppl` of package *spatstat* in R

**Source:** INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011

**Scope:** Reunion Island

---

```
#base.ppp created above is used again

# bw.frac test for optimum bandwidth
bw_frac<- spatstat::bw.frac(base.ppp)
plot(bw_frac, main = "bw.frac")
```

---

The result is:

---

```
bw_frac
## [1] 19747.02
```

---

Using the default settings, the proposed value for the  $\sigma$  value is 19,747 metres.

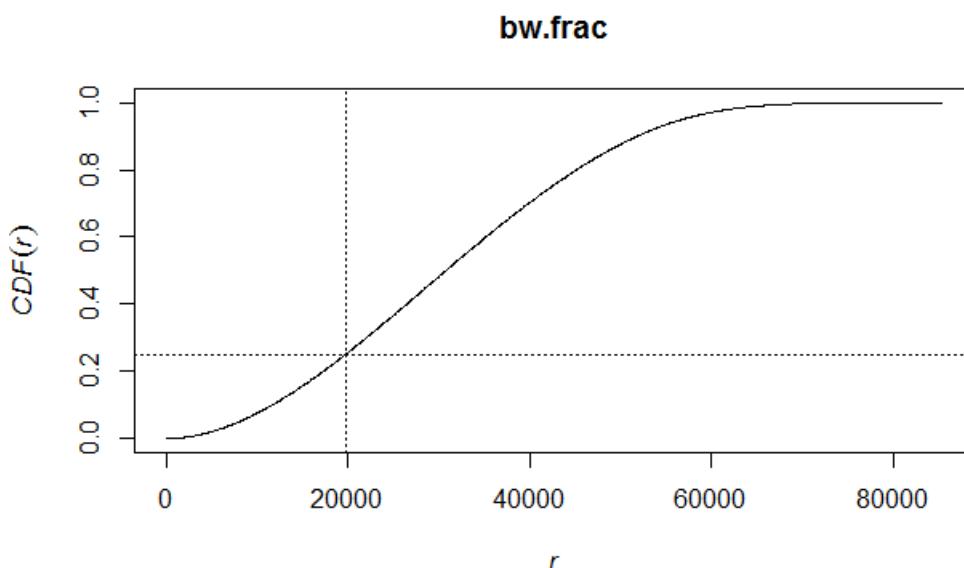


Figure 8.14 – The cumulative  $CDF(r)$  distribution function obtained by function `bw.frac` of package `spatstat` in R

**Source:** INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing Tax (TH) as at 1 January 2011

**Scope:** Reunion Island

#### With function `bw.scott` of `spatstat`

This function is based on the "Scott rule" (see Scott 1992). The idea consist in assuming that the sample is distributed according to a normal law. In this case, a bandwidth estimator is derived, minimizing an error called "mean integrated squared error". The estimator formula includes in particular the sample's standard deviation.

The result is a vector composed of two values — the bandwidths suggested in the direction of the x and y values.

---

```
#the base.ppp created above is used again

#bw.scott test for optimal bandwidth
bw_scott<- spatstat::bw.scott(base.ppp)
```

---

The result is:

---

```
bw_scott
## [1] 2973.548 3455.256
```

---

With the default parameters, the proposed value is the couple (2974; 3455): 2,974 metres in the direction of the x and 3,455 in the direction of y values. According to the package documentation, the value suggested by this test is generally higher than that provided by bw.diggle.

### Summary of results found

Function	$\sigma$ (in metres)
bw.diggle	142
bw.ppl	286
bw.frac	19747
bw.scott	2974 (x) et 3455 (y)

Table 8.1 – "Optimal" bandwidths ( $h = 2\sigma$ ) found using the functions of package *spatstat*

**Source:** INSEE, Localised Tax Revenues System (RFL) as at 31 December 2010 and Housing

Tax (TH) as at 1 January 2011

**Scope:** Reunion Island

In this example, the results sometimes vary widely depending on the methods. The disparities are exacerbated by the unusual distribution of the population on Reunion Island, almost exclusively located on the coast.

### Conclusion

Behind the aesthetic quality of the smoothed maps, however, lies a major trap. By construction, smoothing methods mitigate breakdowns and borders and induce continuous representation of geographical phenomena. The smoothed maps therefore show the spatial autocorrelation locally. Two points close to the smoothing radius have mechanically comparable characteristics in this type of analysis. As a result, there is little point in drawing conclusions from a smoothed map of geographical phenomena whose spatial scale is of the order of the smoothing radius. Intuitively, this amounts to commenting on the homogeneity observed within the spatial units of a choropleth map. In other words, the smoothing radius (the bandwidth) implicitly defines a minimum level of information restitution. As a corollary to these remarks, it is essential to comment only on the phenomena whose order of magnitude is much higher than the smoothing radius.

**References - Chapter 8**

- Baddeley, A. et al. (2015a). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press.
- Brunsdon, C. et al. (2002). « Geographically weighted summary statistics : a framework for localised exploratory data analysis ». *Computers, Environment and Urban Systems*.
- Diggle, Peter J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press.
- Floch, J.M. (2012b). « Détection des disparités socio-économiques - L'apport de la statistique spatiale ». *Documents de Travail INSEE*.
- Palsky, Gilles (1991). « La cartographie statistique de la population au XIXe siècle ». *Espace, populations, sociétés* 9.3, pp. 451–458.
- Scott, D.W. (1992). *Multivariate Density Estimation : Theory, Practice, and Visualization*. New York, Chichester : Wiley.



# 9. Geographically Weighted Regression

MARIE-PIERRE DE BELLEFON, JEAN-MICHEL FLOCH  
INSEE

---

<b>9.1</b>	<b>Why use geographically weighted regression?</b>	<b>232</b>
<b>9.2</b>	<b>Geographically Weighted Regression</b>	<b>234</b>
9.2.1	A model with variable coefficients . . . . .	234
9.2.2	How to estimate the model? . . . . .	234
9.2.3	Choosing the estimation parameters . . . . .	236
<b>9.3</b>	<b>Robust Geographically Weighted Regression</b>	<b>240</b>
<b>9.4</b>	<b>Quality of estimates</b>	<b>246</b>
9.4.1	Accuracy of parameter estimation . . . . .	246
9.4.2	Testing non-stationarity of coefficients . . . . .	247
<b>9.5</b>	<b>A predictive application</b>	<b>247</b>
9.5.1	Problem overview . . . . .	247
9.5.2	Results . . . . .	249
<b>9.6</b>	<b>Precautions to take</b>	<b>250</b>
9.6.1	Multicollinearity and correlation between coefficients . . . . .	250
9.6.2	Interpreting the parameters . . . . .	252

---

## Abstract

Geographically Weighted Regression (GWR) was developed in response to the finding that a regression model estimated over the entire area of interest may not adequately address local variations. The fairly simple principle on which it is based consists on estimating local models by least squares, each observation being weighted by a decreasing function of its distance to the estimation point. Combining these local models makes it possible to build a global model with specific properties. GWR can be used, in particular with the help of associated cartographic representations, to identify where local coefficients deviate the most from the overall coefficients, to build tests to assess whether the phenomenon is non-stationary and to characterise non-stationary. The method is presented using the example of a hedonic pricing model – prices of existing housing in Lyon. We show how to optimally determine the radius of the disk on which local regressions will be performed and we present the estimation results, the robust estimation methods and the tests of coefficients' non stationarity. In addition to this descriptive use, we present a more predictive approach, showing how taking non-stationarity into account makes it possible to improve an estimator over a spatial area. The example is based on a model linking the poor population and the number of beneficiaries of supplementary universal health coverage (CMU-C) in Rennes.

**R** Prior reading of chapter 3: "Spatial autocorrelation indices" is recommended.

## 9.1 Why use geographically weighted regression?

To identify the nature of relationships between variables, linear regression models the dependent variable  $y$  as a linear function of explanatory variables  $x_1, \dots, x_p$ . If you have  $n$  observations, the model is written:

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ik} + \varepsilon_i,$$

where  $\beta_0, \beta_1, \dots, \beta_p$  are the parameters and  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  are the error terms. In this model, the coefficients  $\beta_k$  are considered identical across the study area. However, the hypothesis of spatial uniformity of the effect of explanatory variables on the dependent variable is often unrealistic (Brunsdon et al. 1996). If the parameters vary significantly in space, a global estimator will hide the geographical richness of the phenomenon.

**Spatial heterogeneity corresponds to this spatial variability in the model's parameters or its functional form.** When the territory of interest is well-known, it is often treated in empirical literature by adding dummy variables of geographical zones in the model – possibly crossed with each explanatory variable – by estimating the model for different zones or by conducting tests of geographical stability on the parameters (known as Chow tests). When the number of these geographic areas increases, this treatment nevertheless decreases the number of degrees of freedom and therefore the accuracy of the estimators.

Local regressions can also be used, the spatial application of which is referred to as GWR, Geographically Weighted Regression (Brunsdon et al. 1996). Through the example of the study of property prices in Lyon, we show the interest of performing a geographic regression (example 9.1) and how to implement it (example 9.2).

More complex methods coming from geographical researchers have been developed (Le Gallo 2004), but they remain largely descriptive and exploratory – in particular through graphical representations – as their theoretical behaviour is not fully known, in particular their convergence and their handling of geographic break.

■ **Example 9.1 — Use of a hedonic model to study real estate prices in Lyon.** Mapping changes in real estate prices makes it possible to generally deduct that prices tend to be higher in the centre than in the outskirts (Figure 9.1). However, these high prices may be explained by better quality in the housing which is sold in the centre. The hedonic model is aimed at **isolating the effect of localisation on prices**. The principle of this method is that the price of a property is a combination of the prices of its various attributes

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \tag{9.1}$$

with  $x_{ik}$  the characteristic  $k$  of property  $i$ ,  $\beta_k$  the coefficient associated with this characteristic and  $p$  the number of explanatory variables.

The assumptions underlying the hedonic model are that sellers and buyers are individual agents, without market power, and that this is a situation of perfect competition. The hedonic regression coefficient corresponding to a characteristic informs about the value which the purchasers **at equilibrium at a given time** would give to **an increase in the quantity of this characteristic**.

Figure 9.2 depicts the residuals of a hedonic regression of flats' prices on their physical characteristics. These residuals are not randomly distributed in space – the null hypothesis of the Moran test is rejected. The Moran's I of the distribution of residuals is positive, which is a

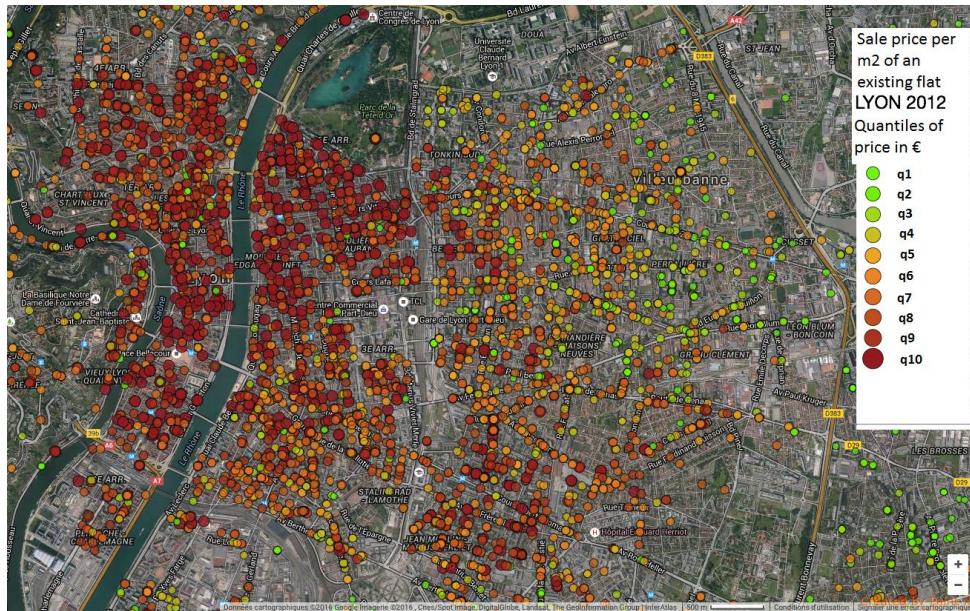
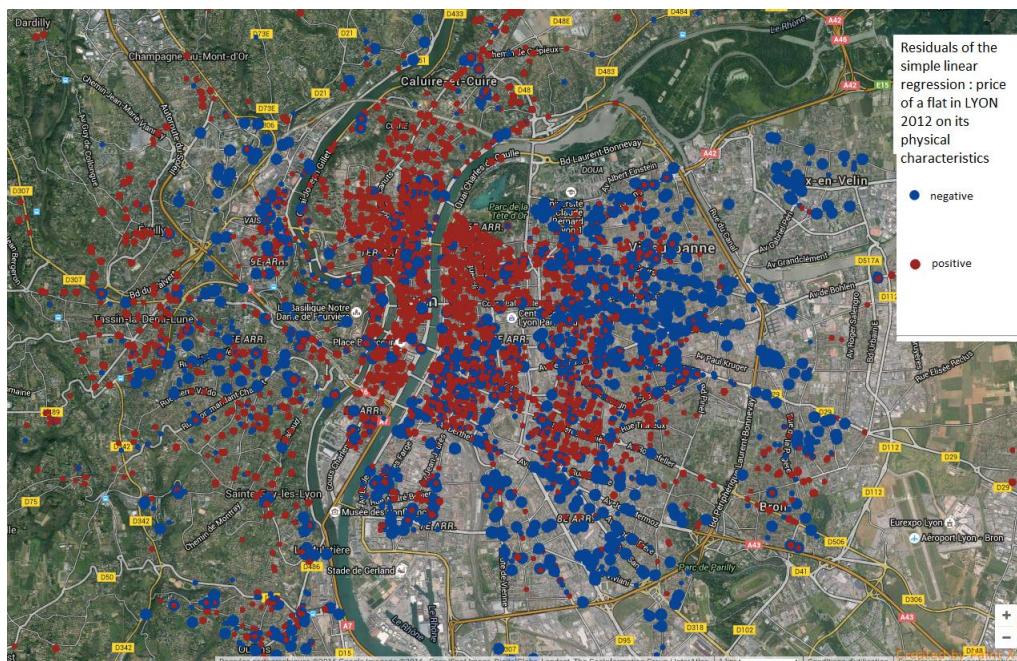
Figure 9.1 – Sale price per m<sup>2</sup> of an existing flat - 2012**Source:** PERVAL base**Scope:** Lyon conurbation

Figure 9.2 – Residuals from the hedonic regression of price on the characteristics of the property

**Source:** PERVAL base**Scope:** Lyon conurbation

sign of positive spatial correlation in the residuals. **The hypothesis of spatial stationarity of the relationship between price and characteristic of the property is not valid.** We can therefore conclude in the existence of **spatial heterogeneity**.

As shown above, in order to take into account the variation in the model parameters with the location, a commonly used method consists in introducing geographical dummy variables as explanatory parameters. Let us examine the evolution of the influence on the price per square meter of an existing flat in Lyon in 2012 of a construction year between 1992 and 2000 as opposed to between 1948 and 1969, depending on the arrondissement in which the property is located.

Arrondissement	Parameter estimate	Significance
1st	1,1511	.
2nd	1,1499	.
3rd	<b>1,1481</b>	***
4th	1,360	**
5th	1,4909	***
6th	<b>1,3085</b>	***
7th	1,1897	***
8th	1,1487	***
9th	1,1981	***

Table 9.1 – Significance of regression coefficients associated with the time of construction

\*, \*\*, \*\*\* are the significance thresholds at 10, 5 and 1%

**Source:** PERVAL base

**Scope:** Lyon conurbation

The value and significance of the coefficients change with the arrondissements (table 9.1). Buyers therefore value the time of construction differently, depending on their location. However, why would the boundaries that define the changes in model match the administrative boundaries? **Geographically Weighted Regression allows study a model that varies spatially in a continuous way.**

■

## 9.2 Geographically Weighted Regression

### 9.2.1 A model with variable coefficients

Geographically Weighted Regression belongs to the category of models with variable coefficients. The regression coefficients are not fixed, they depend on the geographical coordinates of observations. In other words, **the coefficients of the explanatory parameters form continuous surfaces that are assessed at certain points in space,**

$$y_i = \beta_0(u_i, v_i) + \sum_k^p \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \quad (9.2)$$

where  $(u_i, v_i)$  are the geographical coordinates.

### 9.2.2 How to estimate the model?

To estimate the model, the following hypothesis is used: **the closer two observations are geographically, the more similar the influence of the explanatory variables on the dependent variable, i.e. the closer the coefficients of the explanatory parameters of the regression.**

Therefore, to estimate the model with variable coefficients at point  $i$ , we want to use the fixed-coefficients model and include in the regression only the observations close to  $i$ . However, the more points are included in the sample, the lower the variance, but the higher the bias. The solution is therefore to **reduce the importance of the most remote observations by giving each observation a decreasing weight with the distance to the point of interest.**

The model to be estimated is as follows:

$$\mathbf{Y} = (\boldsymbol{\beta} \otimes \mathbf{X})\mathbf{1} + \boldsymbol{\varepsilon} \quad (9.3)$$

$\mathbf{Y}$ : vector  $n \times 1$  of the dependent variable.

$\mathbf{X}$ : matrix  $n \times (p+1)$  of  $p$  explanatory variables + constant

$\mathbf{1}$ : vector  $(p+1) \times 1$  of 1

The coefficients  $\boldsymbol{\beta}$  of the model can be expressed in matrix form:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_1, v_1) & \dots & \beta_p(u_1, v_1) \\ \beta_0(u_j, v_j) & \dots & \beta_p(u_j, v_j) \\ \beta_0(u_n, v_n) & \dots & \beta_p(u_n, v_n) \end{bmatrix} \quad (9.4)$$

The  $\otimes$  operator multiplies each component of the coefficients matrix  $\boldsymbol{\beta}$  by the corresponding element of matrix  $\mathbf{X}$  which contains the characteristics of the observations.

In order to give a weight to observations decreasing with their distance to the point of interest, an estimate is performed using weighted least squares, the weighting being governed by weight matrix  $\mathbf{W}_{(u_i, v_i)}$ . The parameters governing the construction of this matrix are detailed in section 9.2.3.

In accordance with the principle of weighted least squares, coefficients  $\hat{\beta}(u_i, v_i)$  at the point of geographic coordinates  $(u_i, v_i)$  minimize sum 9.5:

$$\sum_{j=1}^n w_j(i)(y_j - \beta_0(u_i, v_i) - \beta_1(u_i, v_i)x_{j1} - \dots - \beta_p(u_i, v_i)x_{jp})^2 \quad (9.5)$$

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = (\mathbf{X}^T \mathbf{W}_{(u_i, v_i)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_{(u_i, v_i)} \mathbf{Y} \quad (9.6)$$

$\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$ , where  $\mathbf{S}$  is the "hat matrix" defined by Equation 9.7. Let's note  $\mathbf{x}_i^T = (1 \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{ip})$  column  $i$  of the explanatory variable matrix  $\mathbf{X}$ . Then

$$\mathbf{S} = \begin{bmatrix} (\mathbf{x}_1^T \mathbf{X}^T \mathbf{W}_{(u_1, v_1)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_{(u_1, v_1)} \\ \vdots \\ (\mathbf{x}_n^T \mathbf{X}^T \mathbf{W}_{(u_n, v_n)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_{(u_n, v_n)} \end{bmatrix} \quad (9.7)$$

**Reminder: ordinary least squares estimation**

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (9.8)$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (9.9)$$

$\mathbf{Y}$ : vector  $n \times 1$  of the dependent variable.

$\mathbf{X}$ : matrix  $n \times (p+1)$  of the  $p$  explanatory variables + the constant.

### 9.2.3 Choosing the estimation parameters

Matrix  $W_{(u_i, v_i)}$  contains the weight of each observation according to its distance to the point  $i$  of coordinates  $(u_i, v_i)$  (Figure 9.3). We assume that observations close to point  $i$  have more influence over the estimated parameters at place  $i$  than more remote observations. The weight of observations therefore decreases with the distance to the point  $i$ . There are several ways of specifying this decrease. Here we show the main parameters governing the decrease function.

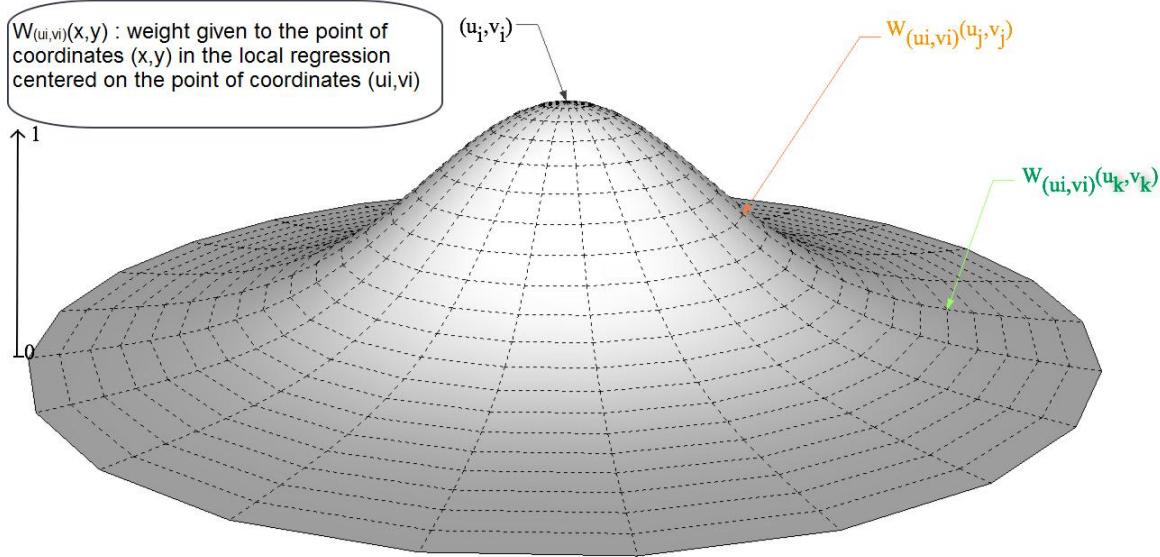


Figure 9.3 – Graphical representation of matrix  $W$

The decrease in the weight of each observation with distance to the point of origin is determined by a **kernel function**. The key parameters of the kernel function are:

- the shape of the kernel ;
- fixed kernel versus adaptive kernel ;
- bandwidth size.

#### Shape of the kernel

We can distinguish the continuous kernel that weights all the observations (Figure 9.4; table 9.2) of the kernel with compact support (Figure 9.5; table 9.3) for which the weight of observations is zero beyond a certain distance. However, **the shape of the kernel only changes the results slightly** (Brunsdon et al. 1998).

Uniform kernel	$w(d_{ij}) = 1$
Gaussian kernel	$w(d_{ij}) = \exp(-\frac{1}{2}(\frac{d_{ij}}{h})^2)$
Exponential kernel	$w(d_{ij}) = \exp(-\frac{1}{2}(\frac{ d_{ij} }{h}))$

Table 9.2 – Continuous kernel

- Choosing a uniform kernel means doing an ordinary least squares regression at each point.
- The Box-Car kernel handles a continuous phenomenon in a discontinuous way.
- Gaussian and exponential kernels weight all the observations, with a weight that tends towards zero with the distance to the estimated point.
- The bisquare and tricube kernels also give observations a decreasing weight with distance, but this weight is zero beyond a certain distance  $h$  called bandwidth).

Box-Car Kernel	$w(d_{ij}) = 1 \text{ if }  d_{ij}  < h, 0 \text{ otherwise}$
Bi-Square	$w(d_{ij}) = (1 - (\frac{ d_{ij} }{h})^2)^2 \text{ if }  d_{ij}  < h, 0 \text{ otherwise}$
Tri-Cube Kernel	$w(d_{ij}) = (1 - (\frac{ d_{ij} }{h})^3)^3 \text{ if }  d_{ij}  < h, 0 \text{ otherwise}$

Table 9.3 – Kernel with compact support

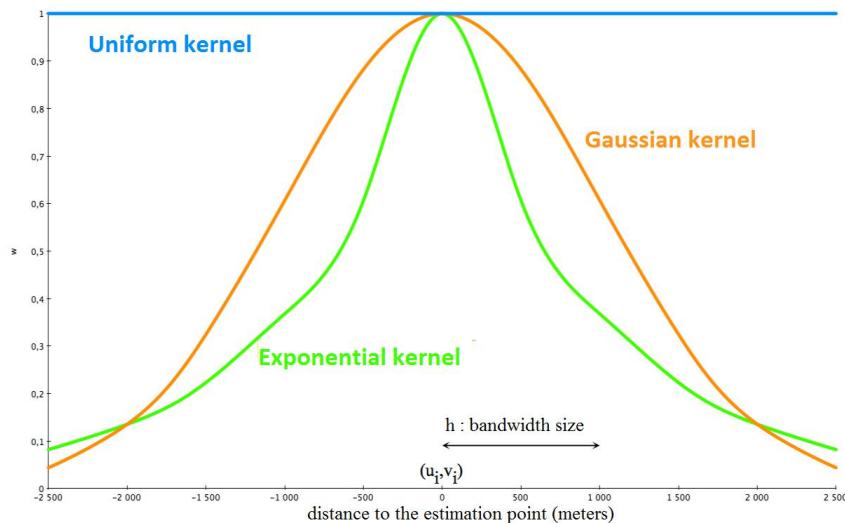


Figure 9.4 – Continuous kernel

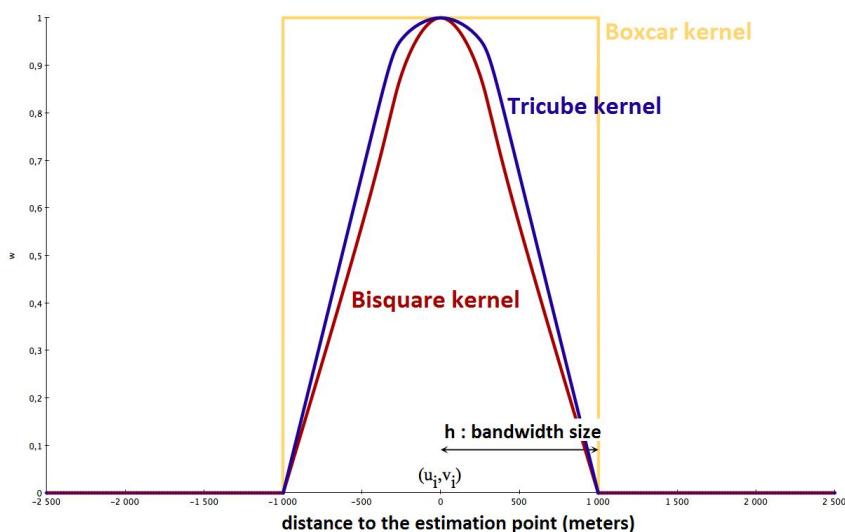


Figure 9.5 – Kernel with compact support

⇒ The bisquare kernel should be preferred in order to optimise calculation time.

### Fixed kernel versus adaptive kernel

**Definition 9.2.1 — Fixed kernel.** The extent of the kernel is determined by the **distance** to the point of interest. The kernel is identical at any point in space (Figure 9.6).

**Definition 9.2.2 — Adaptive kernel.** The extent of the kernel is determined by the **number of neighbours** of the point of interest. The lower the density of the observations, the smaller the kernel (Figure 9.7).

- A fixed kernel is suited to a uniform spatial distribution of data but not very effective in the case of a non-homogeneous distribution. Its radius must be at least equal to the distance between the most isolated point and its first neighbour, which may cause the number of points included in the regression to vary significantly.
- In low-density areas, a fixed kernel that is too small will include too few points in the regression. The variance will be higher.
- In very dense areas, a fixed kernel that is too big will overlook variations on a fine scale. The bias will be higher.

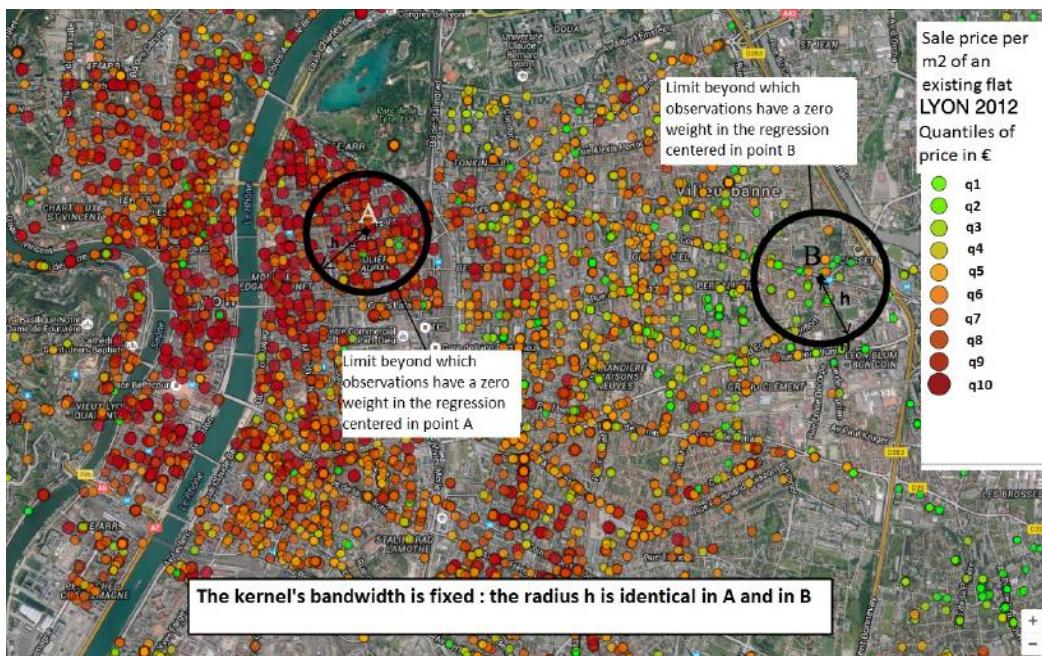


Figure 9.6 – Fixed kernel

Source: PERVAL base

### Definition and choice of bandwidth

The bandwidth is a distance beyond which the weight of the observations is assigned the value 0. **The value of bandwidth  $h$  is the parameter the choice of which has the strongest influence on results.** The larger the bandwidth, the higher the number of observations to which the kernel gives a non-zero weight. The local regression will then include more observations and the results will be smoother than with a small bandwidth. When the bandwidth tends towards infinity, the results of the local regression are similar to those of ordinary least squares regression.

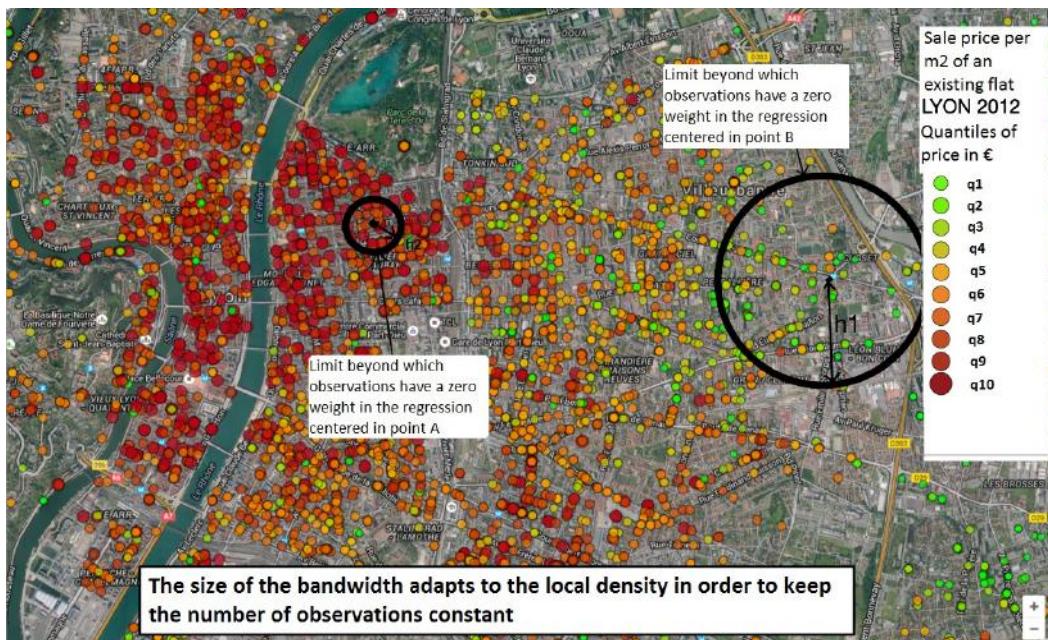


Figure 9.7 – Adaptive kernel

**Source:** PERVAL base

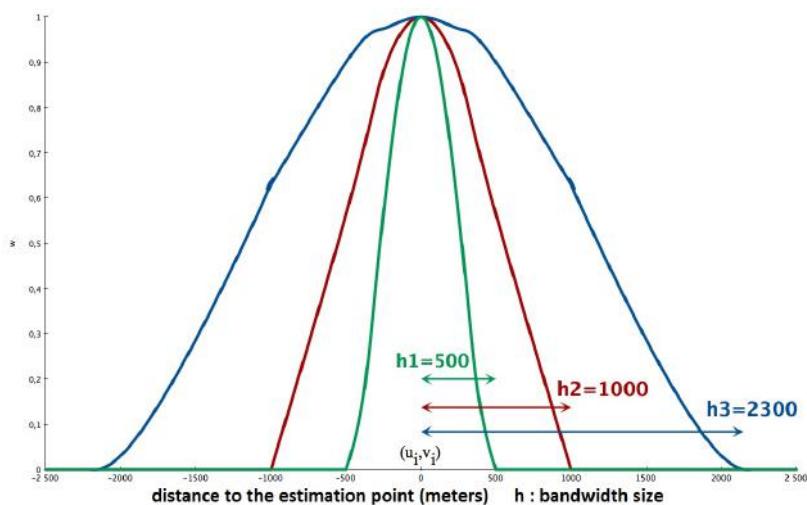


Figure 9.8 – Influence of the choice of bandwidth on the kernel

The choice of the bandwidth is not linked to the model itself, but to the calibration strategy. If the kernel includes points that are too far away, the variance will be low but the bias high. If the kernel only covers the closest points, the bias will be low but the variance high. Several statistical criteria can help choose the most suitable bandwidth. The *GW.model* R package makes it possible to determine the bandwidth that minimises either of the two criteria — the cross-validation criterion and the adjusted Akaike criterion (see boxes 9.2.1 and 9.2.1).

The value of the bandwidth minimising these criteria is also a valuable indication of the relevance of a Geographically Weighted Regression modelling. If the bandwidth tends to the maximum possible – the entire extent of the study area, or all the points – then the local heterogeneity is probably not significant and the GWR is not necessary. Conversely, an extremely small bandwidth should be seen as an alert to the risk that the underlying process could be random (Gollini et al. 2013). It should also be remembered that the bandwidth that minimises the statistical criteria is based on the prediction of the dependent variable, and not of the regression coefficients – which are however those used later to test the validity of the non-stationarity hypothesis.

#### Box 9.2.1 — Cross-validation criteria.

$$CV = \sum_{i=1}^n [y_i - \hat{y}_{\neq i}(h)]^2$$

$\hat{y}_{\neq i}(h)$  is the value of  $y$  at point  $i$  predicted when calibrating the model with all the observations except  $y_i$ . If the model were estimated with all its observations, the optimum bandwidth would indeed be 0 given that, when  $h = 0$  there is no other point than  $y_i$  in the regression; hence  $\hat{y}_i = y_i$  which is the attainable optimum.

Bandwidth  $h$  that minimises CV – the cross-validation score – **maximises the model's predictive power**.

#### Box 9.2.2 — Adjusted Akaike criterion.

$$AIC_c(h) = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n \left\{ \frac{n + tr(S)}{n - 2 - tr(S)} \right\}$$

$n$  is the sample size;  $\hat{\sigma}$  is the estimate of the standard deviation of the error term;  $tr(S)$  is the trace of the projection matrix (hat matrix) of observed variable  $y$  on estimated variable  $\hat{y}$ .

The AIC criterion favours a **compromise between the predictive power of the model and its complexity**. The lower the bandwidth, the more complex the global model. The AIC criterion generally favours larger bandwidths than the CV criterion.

### 9.3 Robust Geographically Weighted Regression

Just like standard linear regression, Geographically Weighted Regression is sensitive to outliers. These points distort local parameters surface estimates (Brunsdon et al. 1996). Since Geographically Weighted Regression takes into account a different model at each point of the space, it is sufficient that one point be unusual **relative to the local context** for the estimate to be distorted. There is, however, more chance for a point to be unusual in relation to the local context rather than the global. By looking for outliers at global level, one thus may overlook points that are unusual locally, but not globally. Two methods have been developed to remedy this problem.

#### Method 1: filter according to standardised residuals

The aim of method 1 is to detect observations with very high residuals and exclude them from the regression.

Let  $e_i = y_i - \hat{y}_i$  be the residual of the estimate at point  $i$ . If  $y_i$  is an outlier,  $e_i$  should have a very high value. However, the residuals do not all have the same variance, so they must be standardised so that they can be compared and a decision made as to which need to be removed from the regression.

Note  $\hat{y} = \mathbf{S}\mathbf{y}$  where  $S$  is the hat matrix defined above.  $\mathbf{e} = \mathbf{y} - \mathbf{S}\mathbf{y} = (\mathbf{I} - \mathbf{S})\mathbf{y}$  with  $\mathbf{e}$  the vector of the residuals.

$var(\mathbf{e}) = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T var(\mathbf{y}) = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T \sigma^2$  with  $\sigma$  the standard deviation of  $y$ . The variances of the  $e_i$ s are therefore the leading diagonal elements on the matrix  $(\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T \sigma^2$ , which in general are not equal (Brunsdon et al. 1996).

Consider  $\mathbf{Q} = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T$  and  $q_{ii}$  the  $i$ th element of the diagonal of  $\mathbf{Q}$

$r_i = \frac{e_i}{\hat{\sigma}\sqrt{q_{ii}}}$  is called the *internally standardised residual*.

If point  $i$  is unusual, including it in the estimate of  $\hat{\sigma}^2$  may produce a bias. The value of  $\sigma$  is thus estimated by excluding the potentially outlier observation  $i$ ,  $\sigma_{-i}$

$r_i^* = \frac{e_i}{\hat{\sigma}_{-i}\sqrt{q_{ii}}}$  is called *externally standardised residual*.

With method 1, the observations for which  $|r_i^*| > 3$  are filtered (the threshold of 3 is proposed by Chatfield 2006).

Disadvantage:  $\mathbf{Q}$  is a matrix  $n * n$  of which **the calculation time is prohibitive for large databases**, at this time and with a machine with conventional calculation power. For example: Brunsdon et al. 1996 deem that this method cannot be used beyond 10 000 observations.

### Method 2: reducing the weight of observations with high residuals

The objective of method 2 is to lower the weight of the observations with high residuals (Huber 1981). After an initial estimation of the model, weight  $w_r(e_i)$  is ascribed to each observation  $i$ . This weight must be multiplied with the weight which varies according to the distance to the point  $i$ . A new matrix  $W$  is thus created, which is the term by term product between the old matrix  $W$  and a matrix  $W_r$  of the residual weights, defined as:

$$w_r(e_i) = \begin{cases} 1 & \text{if } |e_i| \leq 2\hat{\sigma} \\ [1 - (|e_i| - 2)^2]^2 & \text{if } 2\hat{\sigma} < |e_i| < 3\hat{\sigma} \\ 0 & \text{otherwise} \end{cases} \quad (9.10)$$

If none of the residuals from the first regression is higher than two standard deviations, the second model is identical to the first. The observations of which residuals are between two and three standard deviations have their weight reduced in the second regression, while the observations whose residuals exceed three standard deviations are excluded.

### Discussion

Method 2 is much quicker to calculate than method 1 since each cycle only requires the calculation of the  $n$  residuals rather than a matrix  $n * n$ . However, it does not take into account differences in variance between residuals and eliminates more points than does method 1.

## Application with R

Package *GW model* is used to implement Geographically Weighted Regression. The first step consist in calculating the distances between all the observations thanks to function `gw.dist`. Then function `bw.gwr` is used to calculate the bandwidth of the kernel function, optimal with respect to a given statistical criterion. Lastly, the local coefficients of the Geographically Weighted Regression are derived thanks to function `gwr.robust`. The results are contained in an object of class `gwrm`, containing in particular an object of type `SpatialPointsDataFrame`, the contents of which are detailed below.

### Options of function `gw.dist`

- `dp.locat`: coordinates of observations;
- `rp.locat`: coordinates of points at which to calibrate the model (*e.g.* points on a regular grid);
- `p`: governs the choice of distance ( $p=1$ : Manhattan -  $p=2$ : Euclidean);
- `theta`: angle to rotate the coordinates system (useful for Manhattan distance).

### Options the function `bw.gwr`

- `formula`: the model  $y \sim x_1 + x_2 + \dots + x_p$
- `approach`: optimal bandwidth calculation method: CV (Cross Validation) or AIC (Akaike Information Criterion).
- `kernel`: type of kernel: "gaussian", "exponential", "bisquare", "tricube", "boxcar"
- `adaptive`: if TRUE, then the bandwidth is a number of neighbours, and the kernel is adaptive. if FALSE, the bandwidth is a distance, and the kernel is fixed.
- `dMat`: pre-calculated distance matrix.

### Options of function `gwr.robust`

- `regression.points`: geographical coordinates of points where the model will be evaluated.
- `bw`: size of the bandwidth.
- `filtered`: if TRUE, filter the observations according to the value of standardised residuals robust regression method 1) and if FALSE, estimate the model a second time by weighting the observations according to the value of their residuals (robust regression method 2)
- `F123.test`: calculates the Fischer statistics (FALSE by default)
- `maxiter`: maximum number of iterations of the automatic approach (Method 2). It is equal to 20 by default.
- `cut1`:  $\sigma_{cut1}$  is the threshold value of the residuals beyond which the observations have a weight  $< 1$  (set to 2 by default).
- `cut2`:  $\sigma_{cut2}$  is the threshold value of the residuals beyond which the observations have a null weight (set to 3 by default).
- `delta`: tolerance threshold of the iterative algorithm (set to  $1.0e^{-5}$  by default).

### Interpreting the results: the contents of file \$SDF

- The \$SDF file is of "SpatialPointsDataFrame" type, which contains attributes associated with geographic coordinates.
- `c_x`: estimation of the coefficient associated with characteristic x at each point.
- `yhat`: predicted value of y.
- `intrinsic`, `Stud_residual`: residual and standardised residual
- `CV_score`: cross validation score
- `x_SE`: standard error of the estimate of the coefficient associated with characteristic x.
- `x_TV`: t-value of the estimate of the coefficient associated with characteristic x.
- `E_weight`: weight of the observations in the robust regression (to be multiplied by the weight obtained with the kernel function).

■ **Example 9.2 — Application to the study of Lyon real estate prices.** Geographically Weighted Regression makes it possible to study the influence of a property's location on its price, while taking into account spatial heterogeneity — the fact that the influence of the characteristics of a property on its price depends on its location. The coefficient associated with the constant of the geographic regression is the price of a reference apartment — the price of an apartment, once the influence of its physical characteristics has been taken into account.

Meaning of the variables in the example below:

f\_lgpx: logarithm of the price per square metre.  
c\_epoqueA: dummy variable of construction before 1850  
c\_epoqueF: dummy variable of construction between 1981 and 1991  
c\_epoqueG: dummy variable of construction between 1992 and 2000  
c\_mmut1, 2, 3: dummy variable of a transfer in January, February, March, etc.  
c\_sdbn\_2: dummy variable of the existence of two bathrooms.  
c\_cave1: dummy variable of the existence of a cellar.

```
library(GWmodel)
dm.calib <- gw.dist(dp.locat=coordinates(lyon2012))

#Calculation of a distance matrix between the points
bw0 <- bw.gwr(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+
                 c_mmut_3+c_epoqueA+c_epoqueF+c_sdbn_2+c_cave1,
                 data=lyon2012, approach="AIC", kernel="bisquare",
                 adaptive=TRUE,dMat=dm.calib)

gwr.robust.lyon2012 <- gwr.robust(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+
                                         c_mmut_3+c_epoqueA+c_epoqueF+c_sdbn_2+c_cave1,
                                         bw=bw0, kernel="bisquare", filtered=FALSE, adaptive=TRUE,
                                         dMat=dm.calib)

#Extraction of the constant: price of the reference property (Figure 9.9)
lyon2012.intercept.robust <- gwr.robust.lyon2012$SDF[,c(1)]
# 1 is the position of the constant in the file containing the regression
# results.
lyon2012.intercept.robust$Intercept <- exp(lyon2012.intercept.robust$Intercept)

#Extraction of the coefficient linked to building before 1850 rather than
# between 1948 and 1969 (reference period) - Figure 9.10
lyon2012.epoqueA.robust <- gwr.robust.lyon2012$SDF[,c(15)]
lyon2012.epoqueA.robust$c_epoqueA <- exp(lyon2012.intercept.robust$c_epoqueA)

#Estimate of the (non robust) model on a grid of 100**100 metres (Figure
# 9.11)
#Let "quadrillage" be a file of type "SpatialGridDataFrame" covering the
# area to be studied
dm.calib.quadrillage <- coordinates(quadrillage) gw.dist(dp.locat=
    coordinates(lyon2012),rp.locat=coordinates(quadrillage))
gwr.lyon2012<-gwr.basic(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+c_mmut_3+c_
```

```
epoqueA+c_epoqueF+c_sdbn_2+c_cave1,regression.point=quadrillage,bw=bw0,
kernel="bisquare", filtered=FALSE, adaptive=TRUE, dMat=dm.calib.
quadrillage)
```

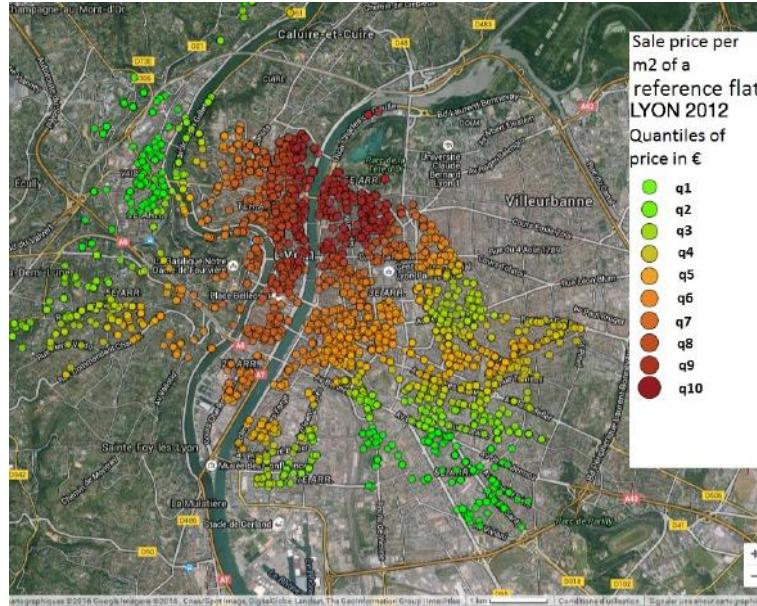


Figure 9.9 – Local constant: price of the reference property

**Source:** PERVAL base

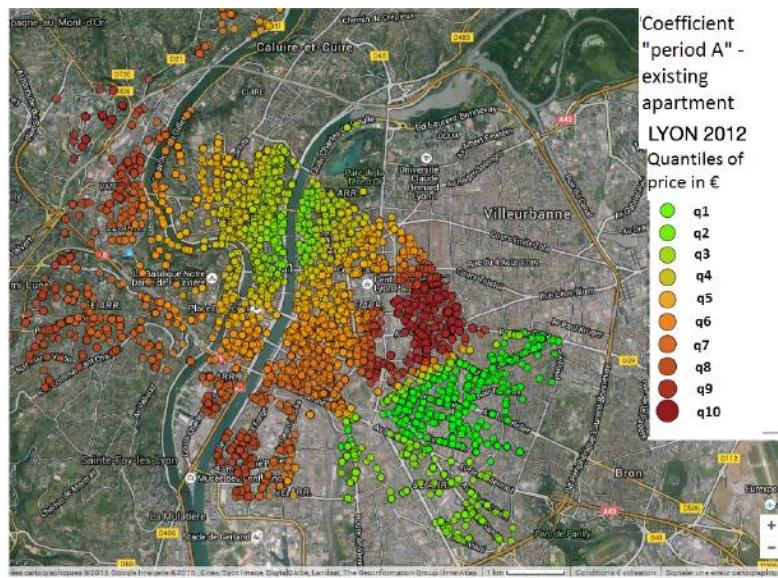


Figure 9.10 – Coefficient associated with building before 1850 rather than between 1948 and 1969 (reference period)

**Source:** PERVAL base

Hedonic regression coefficients vary in space (Table 9.4). Geographically Weighted Regression has made it possible to better understand the spatial richness of changes in the explanatory parameters of real estate prices since the estimates are independent of the arrondissements' administrative boundaries. On Figures 9.9 and 9.10, the points at which the coefficients have been estimated

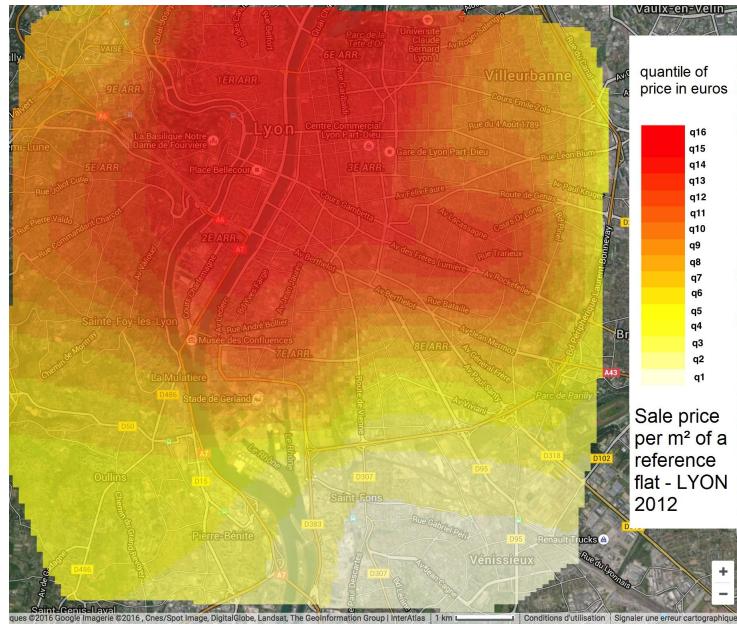


Figure 9.11 – Estimate of real estate prices on a square grid of 100m\*100m

**Source:** PERVAL base

Coefficient	Min	1st quartile	Median	Average	3rd quartile	Max
constant	1666	2220	2668	2705	3088	4030
period A	0.6250	0.9533	1.1480	1.1070	1.2470	1.8190

Table 9.4 – Descriptive statistics of the hedonic GWR parameter estimates

**Source:** PERVAL base

are those where transactions have taken place. However, one of the interests of GWR also lies in the ability to estimate the values of coefficients continuously. Figure 9.11 presents an estimate of parameters on a grid of 100\*100 meters. Section 9.4 shows a method to assess the significance of the spatial variation of parameters. ■

## 9.4 Quality of estimates

### 9.4.1 Accuracy of parameter estimation

When we estimate a GWR with an adaptive kernel in an area where the observations are not very dense, the points used to calibrate the model may have a very low weight – they are located at a long distance from the point of estimation.

Let  $\mathbf{C}$  be the matrix such that:

$$\hat{\beta}(u_i, v_i) = (\mathbf{X}^T \mathbf{W}_{(u_i, v_i)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_{(u_i, v_i)} \mathbf{y} = \mathbf{C} \mathbf{Y} \quad (9.11)$$

The variance of the estimated parameter is:

$$Var[\hat{\beta}(u_i, v_i)] = \mathbf{C} \mathbf{C}^T \sigma^2 \quad (9.12)$$

With  $\sigma^2$  the sum of the standardised residuals of the local regression:

$$\sigma^2 = \sum_i (y_i - \hat{y}_i)^2 / (n - 2v_1 + v_2) \quad (9.13)$$

$$v_1 = \text{tr}(\mathbf{S}) \quad (9.14)$$

$$v_2 = \text{tr}(\mathbf{S}^T \mathbf{S}) \quad (9.15)$$

$$\hat{Y} = \mathbf{S} \mathbf{Y} \quad (9.16)$$

Once the variance of each parameter has been estimated, the standard errors are calculated using Equation 9.17

$$SE(\hat{\beta}(u_i, v_i)) = \sqrt{Var[\hat{\beta}(u_i, v_i)]} \quad (9.17)$$

One can thus calculate confidence intervals for the coefficients.

### Application with R

The \$SDF file containing the results of the Geographically Weighted Regression makes it possible to access the standard errors associated with the various coefficients. For example, in the case of the example of Lyon real estate prices developed previously:

- y: selling price.
- yhat: estimated selling price.
- Intercept\_SE: standard error of the coefficient associated with the constant.
- Intercept\_TV: variation rate of the coefficient associated with the constant.

### 9.4.2 Testing non-stationarity of coefficients

The GWR does not take into account the hypothesis that the coefficients are stationary in a certain geographic area. To verify the relevance of the model, it is interesting to test the non-stationarity of the coefficients. **Do the coefficients vary enough in space to reject the hypothesis that they are constant throughout the study surface?**

In statistical terms, the question can be stated:

- $H_0: \forall k, \beta_k(u_1, v_1) = \beta_k(u_2, v_2) = \dots = \beta_k(u_n, v_n)$
- $H_1: \exists k, \text{ all } \beta_k(u_i, v_i) \text{ are not equal.}$

To answer this question, a simulation method of the Monte Carlo simulation type can be used.

Principle: If there were no underlying spatial phenomena, the geographical coordinates of the observations could be permuted randomly, and the variance would remain unchanged. In a Monte Carlo simulation, the geographic coordinates of the observations can be permuted  $n$  times. This results in  $n$  estimates of the spatial variance of coefficients. The next step consist in estimating the  $p$ -value of the coefficients' spatial variability and rejecting - or maintaining - the null hypothesis that they are stable in space.

It should be remembered, however, that the methods simulating a spatial distribution of the observations depend upon the initial dataset. Leung et al. 2000 describe a more robust and less time-consuming calculation method for testing the coefficients' non-stationarity.

### Application with R

**Function** montecarlo.gwr

- same parameters as function gwr.robust
- nsims: number of simulations
- sortie: vector containing p-values of all GWR parameters

## 9.5 A predictive application

Geographically Weighted Regression has been used primarily to highlight spatial heterogeneity. Like other regression methods, it can also be used for predictive purposes, for example to allocate values to unsampled units in a survey. This section of the article is based on work carried out by E.Lesage and J-M. Floch for the 2015 JMS<sup>1</sup>, and presented at the 2016 workshop dedicated to Advanced Methods for the Analysis of Complex Samplings. In the small area estimation methods, approaches based on models using BLUP (Best Linear Unbiased Predictors) estimators (Chambers et al. 2012) are more and more frequent. The values of unsampled units are replaced by the predicted values from a model whose parameters are estimated using the values of the sampled units. An extension of these methods has been proposed (Chandra et al. 2012) in a non-stationary framework, using geographically weighted regression. The use of Geographically Weighted Regression in small-area estimation methods appears to be favoured in recent literature over methods derived from spatial econometrics, notably using spatial autoregressive models (SAR). Geographically Weighted Regression provides a more flexible way of taking into account phenomena of spatial variability. This consideration of spatial heterogeneity must theoretically improve the accuracy of estimators.

### 9.5.1 Problem overview

At INSEE, empirical research has used GWR to build estimators of population census data on priority neighbourhoods. In these neighbourhoods, 40% of housing is surveyed (over a five-year

---

1. Journée de Méthodologie Statistique – Workshop of Statistical Methodology, organised by INSEE every three years

period), but the sampling design is not optimal, because belonging to a priority neighbourhood is not one of the balancing variables. As there is a high demand for precise information on these neighbourhoods, we have sought to mobilise comprehensive or nearly-comprehensive administrative sources (tax data, health insurance data) to improve the accuracy of the estimators. To do so, a model was estimated over the housing units of the population census (RP), in which the variable of interest was a variable of the census, and the auxiliary variables were variables derived from administrative sources, well correlated to the variable of interest. The estimators made it possible to predict a value for the housing units that had not been sampled. **The estimator of the total of the variable of interest is the sum of the values observed for the sampled units and of the predicted values for the non-sampled units, the sampling weight no longer being taken into account in this calculation.**

This empirical work used Geographically Weighted Regression to model and take into account the significant heterogeneity found in urban data. However, the gain in precision compared to a non-spatial model had not been studied. This is why a comparison of three estimators is proposed from an experimental system based on actual administrative data — the Filosofi source, which makes it possible to calculate the population of low-income households, and the CNAM source (data from the National health insurance fund), which provides the number of CMUC beneficiaries (Universal Supplementary Health Coverage). The Filosofi source is almost comprehensive and provides access to "real" figures on the number of people with incomes below the poverty rate.

Both sources are localised and can theoretically be matched based on their geographic coordinates. For reasons of confidentiality, it was not possible to do so, and we calculated the number of low-income individuals and the number of CMUC beneficiaries on a grid made up of 100 m \* 100 m squares, a compromise with the use of individual data deemed acceptable. These 100m squares play the role of statistical individuals on which we will carry out measurements.

The territory of interest is the municipality of Rennes. Inside the database, a sample of 40% of the squares is drawn, such as what is done in the population census. These squares will serve as a basis for estimating the number of low-income people (Figure 9.12). We have all the information, but for the model, low incomes are only known for sampled squares, while the beneficiaries of the CMUC are known for all squares. We select a sample of size  $n = 856$ , referred to as  $s$ , by simple random sampling without replacement (less complex than the sampling made for the population census). The sampling rate is  $n/N = 40\%$ . In addition, we note  $r$  the complement of sample  $s$  in  $U$  (the set of the inhabited squares in the Rennes region). The calculations made at square level allow calculations on IRIS level, with each square being assigned to an IRIS<sup>2</sup>.

There is a strong linear link between the number of people with low incomes and the number of CMUC beneficiaries. The intercepts vary little from one square to the other. The slopes vary significantly from 1.6 to 3.3. The gradient of local situations is depicted in Figure 9.13. In the approach referred to as "model based", the values  $y_i$  of the non-sampled squares are predicted using the model estimated with all the sampled data and with the auxiliary information  $x$  available for non-sampled squares. We build three estimators,  $j$  representing the IRIS:

**Definition 9.5.1 — Horvitz-Thompson Estimator.**

$$\hat{t}_y(j) = \frac{N}{n} \sum_{i \in s_j} y_i \quad (9.18)$$

2. IRIS are the smallest French administrative delineations

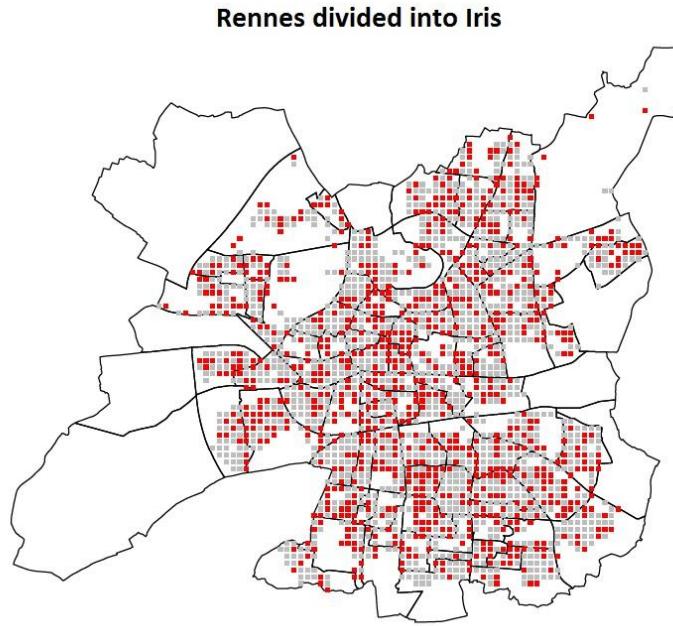


Figure 9.12 – 100 m squares in Rennes, sampled (in red) or not (in gray)

**Definition 9.5.2 — Estimator based on “classic” regression, without taking spatial heterogeneity into account.**

$$\hat{t}_{y,reg}(j) = \sum_{i \in s_j} y_i + \sum_{l \in r_l} \tilde{y}_l \quad (9.19)$$

where  $\tilde{y}_l = \beta^T x_l$

**Definition 9.5.3 — The estimator based on Geographically Weighted Regression.**

$$\hat{t}_{y,RGP}(j) = \sum_{i \in s_j} y_i + \sum_{l \in r_l} \check{y}_l \quad (9.20)$$

where  $\check{y}_l = \hat{\beta}_l^T x_l$  and  $\hat{\beta}_l$  is the vector of the coefficients of the Geographically Weighted Regression for square  $l$ .

## 9.5.2 Results

This process is repeated  $K = 1000$  times. For each IRIS, 1 000 values are derived for each of the three estimators. From these 1 000 values, Monte Carlo estimates of the biases and the root mean squared errors of estimators are developed.

If we note  $\hat{t}_y(j)^{(k)}$  the estimator of the total of variable  $y$  for IRIS  $j$  and for simulation  $k$ , the “Monte Carlo” root mean squared error can be calculated with the following equation:

$$EQM(\hat{t}_y(j)) = K^{-1} \sum_{k=1}^K (\hat{t}_y(j)^{(k)} - t_y(j))^2 \quad (9.21)$$

as the exact total  $t_y(j)$  is known.

The indicator deduced will be used to compare the results of the three estimates, the square

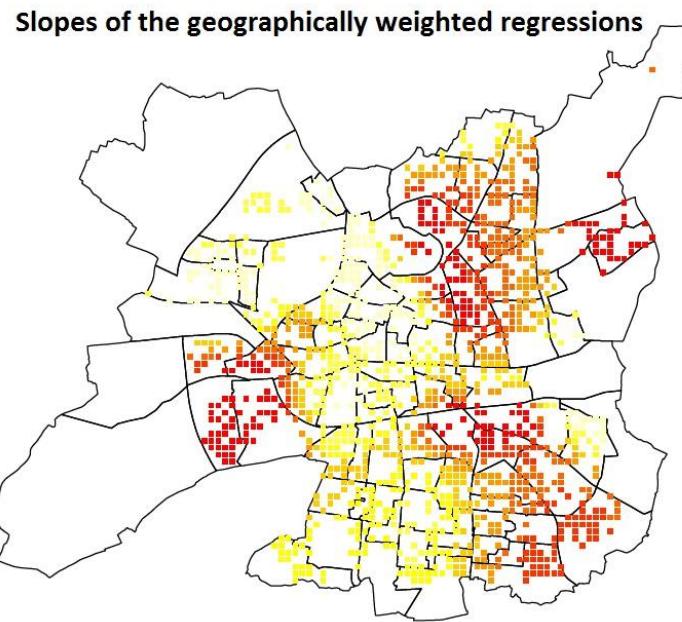


Figure 9.13 – Graphic representation of the slopes of the Geographically Weighted Regressions

**Note:** The scale used is a “heat scale” which ranges from the colour yellow (the highest values) to the colour red (the lowest values).

root of the mean squared error:

$$RCEQMR(\hat{t}_y(j)) = \frac{\sqrt{EQM(\hat{t}_y(j))}}{t_y(j)} \quad (9.22)$$

The IRISes of the municipality of Rennes are ranked in order of increasing population size and the RCEQMs are represented on Figure 9.14 for each of the IRIS.

The first message is the improvement of the accuracy in both model-based approaches, due to the good linear relationship between variable  $y$  (low-income individuals) and variable  $x$  (the CMUC beneficiaries). The RCEQMR is approximately 0.4 for the Horwitz-Thompson estimator, in the order of 0.12 for the models. The difference between the regression and the GWR is not very visible on Figure 9.14. The results are very close. The box-plots of Figure 9.15 make it possible to take the comparison slightly further.

In view of Figure 9.14 the GWR estimator nevertheless proves better than the regression estimator — for 75% of the IRIS, the RCEQMR of the GWR estimator is less than 0.156, the corresponding value for the regression estimator being 0.178.

## 9.6 Precautions to take

### 9.6.1 Multicollinearity and correlation between coefficients

#### Detecting collinearity

In order to estimate the numerous coefficients of a Geographically Weighted Regression, the weighted least squares technique imposes many constraints on the regression parameters (Leung et al. 2000). These constraints can link the GWR coefficients and create multicollinearity problems. The multicollinearity between the coefficients may be responsible for great instability in the coefficients (change of sign when adding a new variable into regression), the counter-intuitive sign of

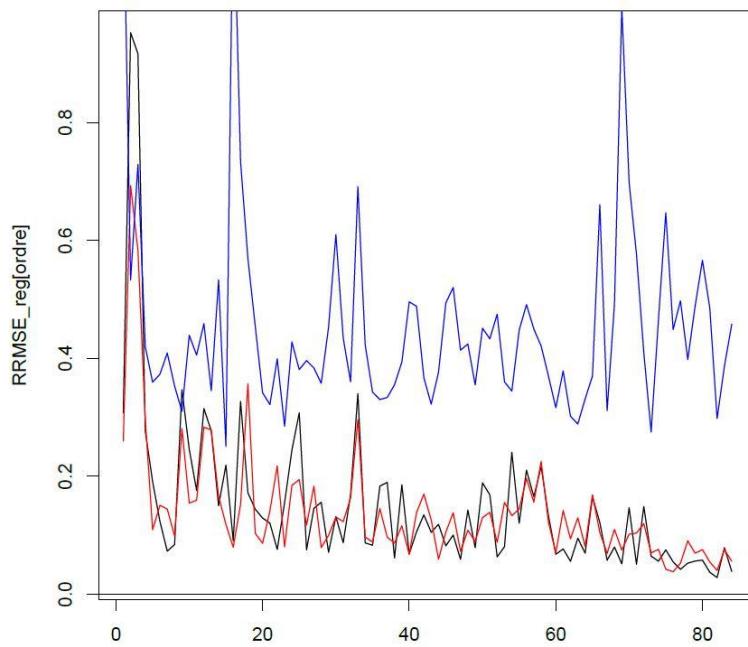


Figure 9.14 – RCEQMR (RRMSE on the figure) of the Horwitz-Thompson estimator (in blue), of the regression estimator (in black) and the estimator by GWR (in red), according to the IRISes, ranked by increasing size

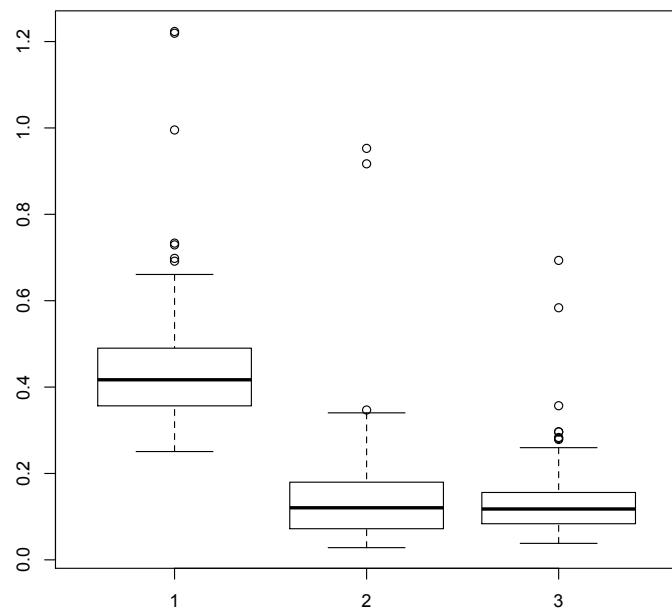


Figure 9.15 – Box-plot of the RCEQMR of the Horwitz-Thompson Estimator (1), of the regression estimator(2) and of the estimator by GWR (3)

one of the coefficients of regression, or high standard errors of the parameters (Wheeler et al. 2005). If the data correlation structure is heterogeneous in space, some regions may show collinearity between their variables, while others will not.

Function `gwr.colin.diag` of package *GW model* allows to implement several types of collinearity detection, in particular local correlations between pairs of coefficients and the variance inflation factors (VIF) for each coefficient. These elements are detailed in Gollini et al. 2013 where examples of application with R are presented.

**Box 9.6.1 — Variance inflation factor: VIF.** Let  $R_j^2$  be the coefficient of determination of the regression of variable  $X_j$  on the  $p - 1$  other variables.

$$VIF_j = \frac{1}{1 - R_j^2}$$

If  $R_j^2$  tends towards 1,  $VIF_j$  tends toward  $+\infty$  hence the term "variance inflation". In general, the literature considers that there is a problem of multicollinearity when a VIF is greater than 10, or when the average of the VIFs is greater than 2 (Chatterjee et al. 2015).

### Taking collinearity into account

One method for reducing the collinearity problems implemented in package *GW Model* is ridge regression. The principle is to increase the weight of the diagonal elements of the variance-covariance matrix to reduce the weight of the non-diagonal elements (which contain the terms of collinearity). In the general case, it can be written that:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y \quad (9.23)$$

The disadvantage of this method is that  $\hat{\beta}$  is biased and standard errors are no longer available. In the case of Geographically Weighted Regression, a local ridge regression can be defined, such that:

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X + \lambda I(u_i, v_i))^{-1} X^T W(u_i, v_i) Y \quad (9.24)$$

$\lambda I(u_i, v_i)$  is the value of  $\lambda$  at location  $(u_i, v_i)$ . It is also possible to use a statistical criterion such as the cross validation score to choose the bandwidth of the local ridge regression.

## 9.6.2 Interpreting the parameters

### The multiple testing problem

When estimating a Geographically Weighted Regression, the result is, at each point, an evaluation of the significance of each coefficient thanks to the calculated t-values. For each coefficient, there are as many t-values as points at which they were estimated. We then come up against the problem of multiple testing presented in Chapter 3, in the case of local spatial autocorrelation indicators.

If we estimate the significance of a coefficient in 100 locations with a significance threshold defined at 95%, we expect to estimate the coefficient as significant in at least 5 locations, simply because of the statistical principle of the test, regardless of any actual correlation between the dependent variable and the explanatory variables. To remedy this problem, a Bonferroni adjustment method can be used, which will increase the value of the threshold beyond which the result of the local test will be judged non-significant - at a constant global significance level. However, adjustment methods have the disadvantage of being often too restrictive, which may lead to some coefficients being judged non-significant when they are actually significant.

Brunsdon et al. 1998 advise caution when using the t-values produced when estimating a GWR. They consider that an area with a large proportion of locally very different coefficients is a better indicator of local non-stationarity than a surface where only a small proportion of coefficients exceeds a significant value.

### **Effect of the local context or incorrect specification**

Before interpreting local coefficient values as characteristics of the local context, it is important to explore the possibility of a poor model specification. For example, the fact that the influence of having a garage on a property price depends on the location may be due to the fact that the density of public parking spaces varies in space, or that the hedonic model is poorly specified.

### **Interpreting the local constant**

In a Geographically Weighted Regression, the constant can vary locally. It may therefore capture all the explanatory power of the exogenous variables, particularly when they have a much more marked influence in certain locations (phenomenon of spatial clustering). In this case, the explanatory variables will appear to be non-significant. If such a phenomenon is suspected, a so-called “mixed” Geographically Weighted Regression can be used, in which the constant does not vary.

## **Conclusion**

GWR which was first suggested in 1998 (Brunsdon et al. 1998) has been the subject of numerous practical applications, in particular in geographic and epidemiological studies. The theoretical foundations have been significantly developed. If some authors have highlighted certain limitations of the method, particularly collinearity problems (Wheeler et al. 2005, Griffith 2008), GWR is now an integral part of spatial analysis tools. It is presented in general works (Waller et al. 2004, Schabenberger et al. 2017, Lloyd 2010, Fischer et al. 2009) as well as in spatial econometrics textbooks (Arbia 2014). Extensions to the method – generalised linear models – have also been proposed.

GWR can be used in two different ways. First of all, it can be used as an exploratory method to detect areas where specific spatial phenomena occur and subject them to a comprehensive study. Secondly, it can help in building a relevant model — the detection of spatial non-stationarity then becomes symptomatic of a problem in the definition of the global model. Brunsdon et al. 1998 consider that most assertions made at a global level about the spatial relationship between objects deserve to be reviewed locally using GWR to test their validity.

Spatial dependency between the error terms decreases when GWR is used, since spatial autocorrelation is sometimes the result of a non modelled instability in parameters (Le Gallo 2004). In addition, GWR makes it possible to calculate spatial autocorrelation indicators for a variable, conditional on the spatial distribution of other variables, which is not possible with the univariate spatial autocorrelation indicators presented in Chapter 3. We therefore encourage the joint study of spatial dependency - with spatial autocorrelation indicators - and spatial heterogeneity - with Geographically Weighted Regression.

## References - Chapter 9

- Arbia, Giuseppe (2014). *A primer for spatial econometrics: with applications in R*. Springer.
- 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.
- Brunsdon, Chris, Stewart Fotheringham, and Martin Charlton (1998). « Geographically weighted regression ». *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.3, pp. 431–443.
- Chambers, Ray and Robert Clark (2012). *An introduction to model-based survey sampling with applications*. Vol. 37. OUP Oxford.
- Chandra, Hukum, Ray Chambers, and Nicola Salvati (2012). « Small area estimation of proportions in business surveys ». *Journal of Statistical Computation and Simulation* 82.6, pp. 783–795.
- Chatfield, Chris (2006). « Model uncertainty ». *Encyclopedia of Environmetrics*.
- Chatterjee, Samprit and Ali S Hadi (2015). *Regression analysis by example*. John Wiley & Sons.
- Fischer, Manfred M and Arthur Getis (2009). *Handbook of applied spatial analysis: software tools, methods and applications*. Springer Science & Business Media.
- Gollini, Isabella et al. (2013). « GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models ». *arXiv preprint arXiv:1306.0413*.
- Griffith, Daniel A (2008). « Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR) ». *Environment and Planning A* 40.11, pp. 2751–2769.
- Huber, Peter (1981). « J. 1981. Robust Statistics ». *New York: John Wiley*.
- Le Gallo, Julie (2004). « Hétérogénéité spatiale ». *Économie & prévision* 1, pp. 151–172.
- 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.
- Lloyd, Christopher D (2010). *Local models for spatial analysis*. CRC press.
- Schabenberger, Oliver and Carol A Gotway (2017). *Statistical methods for spatial data analysis*. CRC press.
- Waller, Lance A and Carol A Gotway (2004). *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons.
- Wheeler, David and Michael Tiefelsdorf (2005). « Multicollinearity and correlation among local regression coefficients in geographically weighted regression ». *Journal of Geographical Systems* 7.2, pp. 161–187.

# 10. Spatial sampling

FAVRE-MARTINOZ CYRIL, FONTAINE MAËLLE, LE GLEUT RONAN, LOONIS VINCENT  
INSEE

---

<b>10.1 General</b>	<b>256</b>
<b>10.2 Constructing primary sampling units of small area and with a constant number of secondary statistical units</b>	<b>257</b>
10.2.1 Rationale . . . . .	257
10.2.2 Method . . . . .	258
10.2.3 Application . . . . .	259
<b>10.3 How to draw a spatially balanced sample</b>	<b>260</b>
10.3.1 The spatially correlated Poisson method (Grafström 2012) . . . . .	262
10.3.2 The local pivotal method (Grafström et al. 2012) . . . . .	263
10.3.3 The cube method . . . . .	263
10.3.4 Ordered spatial sampling methods . . . . .	265
<b>10.4 Comparing methods</b>	<b>268</b>
10.4.1 The principle . . . . .	268
10.4.2 Results . . . . .	269

---

## Abstract

This chapter focuses on using geographical information for survey sampling. Such information can be used at different stages in the sampling design development process. In most surveys carried out face-to-face, multistage sampling designs are used, so as to lower data collection costs through geographically concentrated interviews. Using a geocoded sampling frame is therefore decisive for constructing the primary sampling units. This geographic information can also be used at selection stage to improve the statistical efficiency of the sample when the variables of interest are positively autocorrelated.

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

## Introduction

Eurostat’s Geostat 2 project (2015-2017) was intended to provide a reference framework within which geocoded statistical information could be produced efficiently and used easily. Regarding design-based surveys, the final project report *A Point-Based Foundation for Statistics*, identifies at least three steps of survey design that could benefit from a geocoded sampling frame. Firstly, upstream, when the collection method is face-to-face, precise knowledge of the location of all statistical units allows the creation of primary sampling units<sup>1</sup> (PSU). Knowing the characteristics of these PSUs makes it easier to manage the interviewers’ network while preserving the statistical qualities of the sampling. Secondly, whatever the data collection method, geographical information makes it possible, given certain conditions, to improve the accuracy of estimates by using spatial sampling methods. Thirdly, during the data collection phase, knowing the location of the statistical units sampled makes it easier to identify them when the quality of the addressing is not sufficient.

This chapter focuses exclusively on the first two points. In the first part, we briefly review the sampling theory framework. The second part then offers a method for constructing the smallest primary sampling units in terms of area while having a constant number of statistical units. The third section is dedicated to presenting different spatial sampling methods, while the last part empirically compares their properties, using simulation.

Among the rich literature on the subject, we rely on or direct the reader to Benedetti et al. 2015.

### 10.1 General

The purpose of the sampling theory is to estimate the value of a parameter  $\theta$  measured on a population  $U$  of size<sup>2</sup>  $N$ . We can think of  $\theta$  as a function of the values taken by one or more variables of interest associated to each statistical units. Let  $y_i$  be the value of the variable  $y$  for the statistical unit  $i$  in  $U$ . The survey statistician does not have access to  $y_i$  except for a sub-part of the population, referred to as the sample and expressed as  $s$ . He or she aggregates the values observed on the sample thanks to a function called the estimator, taking value  $\hat{\theta}(s)$  for  $s$ . Estimating  $\theta$  by  $\hat{\theta}(s)$  is known as statistical inference. Properties of statistical inference are described only if  $s$  is chosen randomly.

A sampling design is a probability law across the set  $\mathcal{P}(U)$  of parts (samples) of  $U$ . The classic notation of a random variable with values in  $\mathcal{P}(U)$  is  $\mathbb{S}$ . A sampling design in which all samples with a size different from  $n$  ( $n \in \mathbb{N}^*$ ) have zero probability of being selected is said to be of fixed size  $n$ . It is generally complex to manipulate a probability law on  $\mathcal{P}(U)$ . This is why the survey statistician works with summary versions of the law of  $\mathbb{S}$ , *i.e.* first-order and second-order inclusion probabilities. They refer respectively to the probability of inclusion of unit  $i$  in the sample or the joint probability of inclusion of units  $i$  and  $j$  in the sample:  $\pi_i = \mathbb{P}(i \in \mathbb{S})$  and  $\pi_{ij} = \mathbb{P}((i, j) \in \mathbb{S})$ .

Estimating  $\theta$  by  $\hat{\theta}$  is subject to multiple errors:

- coverage error: some statistical units in the population cannot be selected since they do not appear in the sampling frame;

---

1. Primary sampling units are a sub-division of the population often based on geographical criteria. The first stage of selection of PSUs, and the second stage of selection of individuals in these PSUs, is such that collection can be concentrated and costs be reduced when the survey is conducted face-to-face.

2. In this chapter, in contrast to the previous chapters, the notion of “size” with respect to a geographical area refers to the number of statistical units present inside, not to the surface.

- non-response error: some individuals have an unknown value of  $y_i$  even when they are selected;
- measurement error: collecting incorrect value  $y_i^*$  instead of  $y_i$ .

An estimator with an expected value different from  $\theta$  is said to be biased, whereas the variability of values  $\hat{\theta}(s)$  is assessed using the variance of  $\hat{\theta}$ . The objective is to make the bias and variance as small as possible, by paying special attention to the conditions in which information is collected and/or by judiciously choosing the sampling design.

Out of all parameters to be estimated, the most traditional is the population total of one variable of interest  $y$ :  $\theta = t_y = \sum_{i \in U} y_i$ . Out of all the different possible estimators of  $t_y$ , we will focus on the Narain-Horvitz-Thompson estimator:  $\hat{t}_y = \sum_{i \in S} y_i / \pi_i$ . In the absence of coverage, non-response and measurement errors, this estimator is unbiased. Its variance for a fixed size design is:

$$V(\hat{t}_y) = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (10.1)$$

By analysing Equation 10.1, one can glean indications on the sampling designs to be used to achieve a more accurate estimate of  $t_y$ . If the  $\pi_i$ s are proportional to the  $y_i$ s, the variance equals zero. This solution being impossible in practice, an alternative involves using  $\pi_i$  proportional to  $x_i$ , where  $x$  is an auxiliary variable known for all statistical units and correlated with  $y$ .

This strategy is valid when the survey is mono-thematic (only one variable of interest  $y$ ). The use of such probabilities for another variable of interest  $y'$  uncorrelated with  $x$  can indeed result in highly imprecise estimates. Therefore, when the survey is multi-thematic, statisticians often prefer choosing equal first-order inclusion probabilities. Equal first-order-inclusion probabilities make it possible to "reduce to a minimum the variances that would emerge in the most unfavourable configurations (referred to as the MINIMAX optic), [...] i.e. for the variables that are most likely to impair the accuracy of the estimates" (Ardilly 2006).

When working with a set of fixed first-order-inclusion probabilities, the design should ascribe large  $\pi_{ij}$  when  $y_i/\pi_i$  is very different from  $y_j/\pi_j$ . In the case of spatialised variables and on the assumption that spatial autocorrelation decreases with distance, distant rather than close statistical units should be selected.

## 10.2 Constructing primary sampling units of small area and with a constant number of secondary statistical units

### 10.2.1 Rationale

Sometimes, organisational constraints imply that the face-to-face collection method is conducted on a low population density area. Then the two-stage sampling method is generally preferred. In order to reduce the costs arising from the interviewers' trips, the first-stage selection will include geographic entities (primary sampling units, PSUs) the geographical extent of which must be as small as possible. To simplify, a PSU selected in this manner is then assigned to one interviewer only. Within each PSU, secondary sampling units (SSUs) are selected, each matching up with a statistical unit to be interviewed (individuals in their main dwelling, companies). In order to ensure sufficient workload for the interviewers for one or more surveys, each PSU must also include a minimum number of secondary sampling units.

For a network consisting of  $m$  interviewers and a final sample of  $n$  secondary sampling units,  $m$  PSUs are selected proportionally to their number of secondary sampling units:  $\pi_i^{(1)} = m(N_i/N)$  for

PSU  $i$  bringing together a total of  $N_i$  secondary sampling units. Assuming that  $m$  divides  $n$ , in each of these  $m$  PSUs,  $n/m$  SSUs are drawn based on an equal-probability design:  $\pi_j^{i(2)} = n/(mN_i)$  for secondary sampling unit  $j$  in PSU  $i$ . The final inclusion probability is constant:  $\pi_j^i = \pi_i^{(1)}\pi_j^{i(2)} = n/N$ .

The PSUs are constructed by combining the finest geographical meshes available in the sampling frame. When these meshes remain coarse, for example municipalities, the number of SSUs in the final PSUs proves a difficult parameter to control. As a consequence, at first-sampling stage, the design does not benefit from the MINIMAX property referred to above, since probabilities are proportional to the number of SSUs. This property can be met if the PSUs are of equal size. Equal-size PSUs may also prove preferable for other reasons connected with coordinating the samples (selecting disjoint or nested samples). Note that:

- the complementary  $\bar{\mathbb{S}} = U \setminus \mathbb{S}$  to an equal-probability sample  $\mathbb{S}$ , is itself an equal-probability sample;
- a random sample  $\mathbb{S}_2$ , selected with equal probabilities in a sample  $\mathbb{S}_1$  itself selected with equal probabilities, is itself a equal-probability sample.

The ideal solution is therefore to construct PSUs covering a small geographical area while having equal numbers of secondary sampling units.

### 10.2.2 Method

The problem of constructing equal-size primary sampling units having a small geographical area is a particular case of the more general problem consisting of constructing classification which is subject to size constraints. This topic has enjoyed renewed interest in the recent literature (Malinen et al. 2014, Ganganath et al. 2014, Tai et al. 2017). The aim is to subdivide the territory into classes within which the dispersion of geographical coordinates is as low as possible, while having an expected number of units per class. Here, we introduce a method initially developed to determine PSUs for the French Labour Force Survey (Loonis 2009), and recently considered among other possibilities to establish PSUs for the French master sample for households surveys (Favre-Martinoz et al. 2017).

The general principle is as follows:

1. the statistical units are geocoded according to the most fine-grained geo-referencing possible. Due to the quality of geo-referencing or the nature of data, the number of statistical units  $n_{xy}$  located at the coordinate point  $(x; y)$  may be greater than 1.
2. A path is drawn through all known locations. For this purpose, the methods discussed in Chapter 2: “Codifying the neighbourhood structure”, are used. Insofar as there is no need for the path to return to its starting point, the Hamilton path has been chosen since it is the shortest (Hamilton path minimises the sum of the distances between two consecutive points without setting a starting or finishing point).
3. To construct  $M$  zones, we go through the whole path from the starting point, cumulating  $n_{xy}$  units along the way. When the total exceeds the threshold  $c \simeq \frac{N}{M}$ , the first PSU is constructed. The process is then repeated, from the first point not yet visited on the path.
4. Under ideal conditions where  $M$  divides  $N$  and  $n_{xy} = 1$  for all pairs  $(x, y)$ , the procedure results in geographically homogeneous primary sampling units of equal size. This heuristic approach does not, however, lead to a global optimum. As in any classification, a consolidation procedure needs to be provided to manage any atypical geographical situation and/or PSU size that is too remote from  $c$ . This type of situation can arise, for example, when the last point on the path integrated into a PSU has a very high  $n_{xy}$  value, or when dealing with the last PSU formed.

In the following section, we implement this procedure. We focus in particular on how to construct the path when the number of secondary sampling units is high.

### 10.2.3 Application

Figure 10.1 shows the results when the general strategy described above is applied to the Alsace region (former region, before the restructuring of the French regions in 2016). For the purposes of the French Labour Force Survey (LFS), the main dwellings were gathered in PSUs of 2,600 main dwellings (Figure 10.1b), divided into sectors with 120 main dwellings (Figure 10.1c).

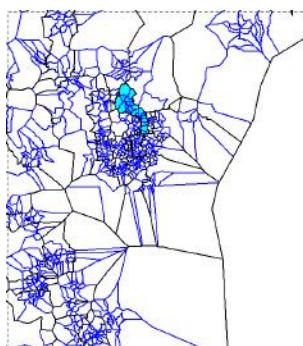
Due to the computation time needed to construct the PSUs, the approximately 616 000 main dwellings were initially grouped into 80 000 grid cells whose resolution is 100 meters (Figure 10.1a), which therefore constitute the most fine-grained georeferencing of statistical units. To construct sectors within the PSUs, the main dwellings are by nature geocoded at building level. The original cells or buildings show high variability in size that implies variability in the  $n_{xy}$  as well. This partly explains the slight variability in the size of the PSUs and sectors (Table 10.1).



(a) 616 000 main dwellings, in 80 000 cells 100 m per side...



(b) ... are gathered in homogeneous PSUs of 2 600 main dwellings...



(c) ... and divided into sections of 120 main dwellings each.

Figure 10.1 – Construction of zones with a small geographical area and an equal number of main dwellings in Alsace

Constructing sectors of 120 main dwellings, from a large number of main dwellings may lead to performance issues. When using a Euclidean distance, the shortest Hamilton path can be computed exactly and easily if the number of points does not exceed a few hundred. When dealing with

Order of Fractile	Size of original cell	Size of PSUs (Figure 10.1b)	Size of sectors (Figure 10.1c)
100%	378	2776	139
99%	59	2757	131
95%	23	2685	130
90%	15	2640	130
75%	9	2606	124
50%	5	2595	119
25%	2	2591	118
10%	1	2587	118
5%	1	2502	117
1%	1	2491	111
0%	1	2479	99

Table 10.1 – Quantiles of the number of main dwellings in the cells, PSUs and sections of Figure 10.1

several thousand points, it is not reasonable nor useful to compute the exact optimal path. We therefore propose an approximate method aiming at constructing a path that fits for the purpose. The different steps in this approximate method are described below and illustrated in Figure 10.2 for one given PSU. This PSU is composed of 2,600 main dwellings and 1,085 buildings.

1. The 1,085 buildings and 2,600 main dwellings of the PSU in blue in Figure 10.1c, are gathered using the k-means method<sup>3</sup> into 20 different but geographically consistent classes. This classification is carried out with the geographical coordinates of the buildings. It should be noted that  $20 \simeq \frac{2600}{120}$  (Figures 10.2a and 10.2b).
2. A **Hamilton** path is drawn through the barycentres of the 20 classes so that they can be ordered (Figure 10.2c).
3. In a given class  $i$ , the **buildings** are sorted according to **two sub-classes** (Figure 10.2d):
  - (a) the first comprises the buildings in class  $i$  that are closer to  $G_{i-1}$  (barycentre of the previous class) than to  $G_{i+1}$  (barycentre of the next class) and increasingly sorted by distance to  $G_{i-1}$ ;
  - (b) the second comprises the buildings in class  $i$  that are closer to  $G_{i+1}$  (barycentre of the next class) than to  $G_{i-1}$  (barycentre of the previous class) and decreasingly sorted by distance to  $G_{i+1}$ ;
4. By construction, the first buildings in class  $i$  are close to the last buildings of class  $i-1$ , and the last buildings in  $i$  are close to the first in class  $i+1$ . Following the path thus means running along buildings by class, by sub-class and finally by increasing or decreasing distance, depending on the case (Figure 10.2e). If necessary, main dwellings inside a building can be sorted by floor.

### 10.3 How to draw a spatially balanced sample

The overall considerations have shown that the more the sampling design selects individuals geographically distant from one another, the more the estimation will be precise for a spatially autocorrelated variable. Grafström et al. 2013, for example, have formalised these considerations more explicitly. In this section, we detail the methods for selecting spatially balanced samples. Existing methods can be grouped into two families.

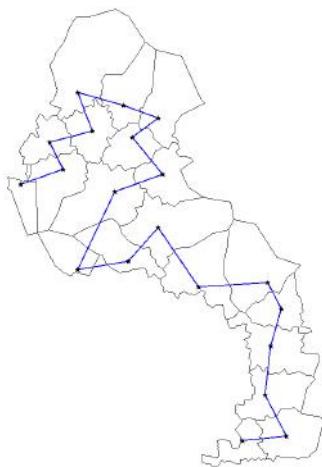
3. The k-means method aims at creating homogeneous classes by maximising between-class variance and minimising within-class variance.



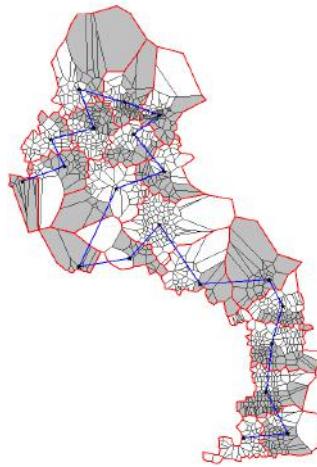
(a) The buildings and their related Voronoï polygons...



(b) ... are grouped, by k-means on the coordinates, into approximately 20 clusters of varying size.



(c) A path through the cluster's barycentres...



(d) ... makes it possible for buildings to be classified according to whether they are closer to the barycentre of the previous cluster (white) or the next one (grey) ...



(e) ... and thus to create a path passing through all the buildings.



(f) Following this path, sectors of 123 to 128 main dwellings are built.

Figure 10.2 – Main dwellings divided into 120-unit sections

Within the first family, the inclusion probabilities are updated locally in order to limit the selection of two neighbouring units. These methods include the spatially correlated Poisson sampling method (Grafström 2012), the local pivotal method (Grafström et al. 2012), and the local cube method (Grafström et al. 2013).

Within the second family, we turn the problem of proximity between units in several dimensions into a problem of order in  $\mathbb{R}$ . Then, the sampling is performed excluding two nearby units, basing on the sorted file. This family of methods includes the *General Randomized Tessellation Stratified* (GRTS, Stevens Jr et al. 2004) method, the method based on a Peano curve (Lister et al. 2009), or on Traveling-Salesman Problem (TSP) algorithm (Dickson et al. 2016).

### 10.3.1 The spatially correlated Poisson method (Grafström 2012)

The spatially correlated Poisson sampling is an extension of the correlated Poisson sampling (*Correlated Poisson Sampling*, CPS) proposed by Bondesson et al. 2008 to perform realtime sampling. The CPS method is based on sequential and orderly sampling of units. Units are ordered with indices ranging from 1 to  $N$ . The decision is first made as to unit 1, then unit 2, up to unit  $N$ . In the case of real time sampling, the order of the indices is a pre-established order of sampleable units. In the case of a spatial sampling, the order can be based on the proximity of the units, in accordance with an Euclidean distance function. At each stage, the inclusion probabilities are updated so as to create a positive or negative correlation between the unit selection indicators.

More specifically, the first unit is included in the sample with probability  $\pi_1^0 = \pi_1$ . If unit 1 was included, we set  $I_1 = 1$ . More generally speaking, at stage  $j$ , unit  $j$  is selected with probability  $\pi_j^{j-1}$  and the inclusion probabilities of units  $i \geq j + 1$  are updated as follows:

$$\pi_i^j = \pi_i^{j-1} - (I_j - \pi_j^{j-1}) w_j^i, \quad (10.2)$$

where  $w_j^i$  is the weight given by unit  $j$  to units with indices  $i \geq j + 1$ . The inclusion probabilities are updated stage by stage, with at most  $N$  stages until the selection indicator vector is obtained.

The choice of weights  $w_j^i$  is crucial as it helps determine whether a positive or negative correlation is introduced between the selection indicators. Bondesson et al. 2008 give the expression of these weights for some conventional sampling designs, and a general expression for any sampling design. Consequently, this method is very general: any sampling design with fixed first-order inclusion probabilities can be implemented by the CPS method. Only the expression and conditions related to weights<sup>4</sup> may vary according to the design. For example, for a fixed-size design, the sum of the weights  $w_j^i$ ,  $j < i$  must be equal to 1. In case of positive spatial auto-correlation (wherein nearby units are similar), the associated weights should be chosen positively, so as to introduce a negative correlation between the sampling selection indicators. It therefore seems appropriate to carry out a global spatial autocorrelation test to determine the sign of the weights to be used in this method.

Grafström 2012 suggests two versions for the weights  $w_j^i$ . Here, we show the version considering a Gaussian distribution. In this case, the weights are defined as:

$$w_j^i \propto \exp(-[d(i, j) / \sigma]^2), \quad i = j + 1, j + 2, \dots, N. \quad (10.3)$$

Since the sum of the weights must be equal to 1, the proportionality constant is set. These weights are all the larger as the units are close to the unit  $j$ . Thus, the closer is unit  $i$  (in the sense of distance  $d(i, j)$ ) to unit  $j$ , the lower is probability  $\pi_i^j$ , and spatially balanced sampling can be carried out. Parameter  $\sigma$  makes it possible to manage the dispersion of these weights, and therefore distribute the update of inclusion probabilities in a wider or smaller neighbourhood, as needed.

---

4. the conditions imposed on weight are linked to the conditions imposed on the inclusion probabilities, *i.e.* the sampling design.

This method is implemented in package *BalancedSampling* in R (Grafström et al. 2016) using the function `scps()`.

### 10.3.2 The local pivotal method (Grafström et al. 2012)

#### Review of the local pivotal method

The local pivotal method is a sampling procedure thanks to which a sample with equal or unequal inclusion probabilities can be selected (Deville et al. 1998). At each stage of the algorithm, the inclusion probabilities of two units  $i$  and  $j$  in competition are updated and at least one of the two units is selected or rejected.

The inclusion probability vector of the two competing units  $(\pi_i, \pi_j)$  is updated according to the following rule (fight between units  $i$  and  $j$ ):

- if  $\pi_i + \pi_j < 1$ , then:

$$(\pi'_i, \pi'_j) = \begin{cases} (0, \pi_i + \pi_j) \text{ with probability } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ with probability } \frac{\pi_i}{\pi_i + \pi_j} \end{cases}$$

- if  $\pi_i + \pi_j \geq 1$ , then:

$$(\pi'_i, \pi'_j) = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ with probability } \frac{(1 - \pi_j)}{(2 - \pi_i - \pi_j)} \\ (\pi_i + \pi_j - 1, 1) \text{ with probability } \frac{(1 - \pi_i)}{(2 - \pi_i - \pi_j)} \end{cases}$$

This procedure is repeated until an inclusion probability vector emerges containing  $N - n$  times the number 0 and  $n$  times the number 1, which will completely determine the selected sample (steps at most  $N$ ).

#### Extension to spatial sampling

The local pivotal method (Grafström et al. 2012) is a spatial extension of the pivotal method. The idea of the method is still to iteratively update the inclusion probabilities vector  $\pi$ , but this time, at each step, we select for the contest two neighbouring units, in terms of a certain distance (*e.g.* a Euclidean distance). Several various methods can be used to select these two neighbouring units:

- **LPM1**: two units as close as possible to one another are selected to participate in the contest, *i.e.* one unit  $i$  is randomly selected among  $N$  population units, then unit  $j$  closest to  $i$  is selected to participate if and only if  $i$  is also the closest unit to  $j$  (at best  $N^2$  steps, at worst  $N^3$  steps);
- **LPM2**: two neighbouring units are selected to participate in the contest, *i.e.* one unit  $i$  is randomly selected from among  $N$  units of the population, then unit  $j$  closest to  $i$  is selected to participate in the fight ( $N^2$  steps);
- **LPM K-D TREE**: the two neighbouring units are selected using spatial partitioning  $k$ - $d$  tree (Lisic 2015) making it possible to search for closer neighbours quicker (complexity of the algorithm in  $N \log(N)$ ).

These three local pivotal methods are implemented in C++ in package *BalancedSampling* in R.

### 10.3.3 The cube method

#### General information about the cube method

Balanced sampling is a procedure aimed at providing a sample that complies with the following two constraints:

- the inclusion probabilities are respected;

- the sample is balanced on  $p$  auxiliary variables. In other words, the Narain-Horvitz-Thompson estimators of the totals of the auxiliary variables are equal to the totals of these auxiliary variables in the population:

$$\sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i \quad (10.4)$$

An algorithm for making such a sampling is called the cube algorithm. To describe the principle, it is appropriate to use the following geometric representation. A sample is one of the vertices of a hypercube with dimension  $N$ , expressed as  $C$ . All  $p$  constraints, recapitulated in Equation 10.4, define a hyperplane with dimension  $N - p$ , expressed as  $Q$ . Using  $K = Q \cap C$ , we depict the intersection between the cube and the hyperplane. A graphical representation of the problem in dimension 3, derived from article Deville et al. 2004, is shown in Figure 10.3.

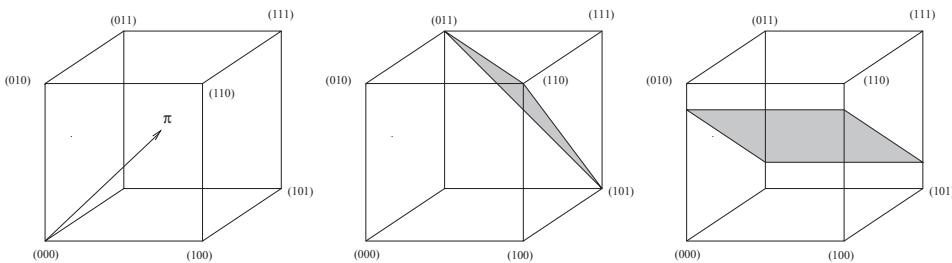


Figure 10.3 – Graphical representation of the cube for  $N = 3$  and different possible configurations of the space subject to constraint, with  $p = 1$

The cube algorithm is divided into two phases. The first phase, referred to as the "flight phase" (Figure 10.4), is a random walk starting with the inclusion probabilities vector and making them change in  $K$ . For this, we start from  $\pi(0) = \pi$ , then update the inclusion probabilities vector by choosing a vector  $u(0)$  such that  $\pi + u(0)$  remains within the space of the constraints. By following the direction indicated by vector  $u(0)$ , we necessarily end up on one face of the cube. The way to update the inclusion probabilities vector is then provided by parameters  $\lambda_1^*(0)$  and  $\lambda_2^*(0)$ , chosen so that updated vector  $\pi(1)$  reaches one face of the cube. The update is chosen randomly so that  $E(\pi(1)) = \pi(0)$ . The operation is then repeated by choosing a new vector  $u(1)$  for the direction and a new direction for updating the inclusion probabilities. This random walk stops when it reaches a vertex point  $\pi^*$  of  $K$ . At the end of this first phase, vertex  $\pi^*$  is not necessarily a vertex of the cube  $C$ . Let  $q$  be the number of non-integer components in vector  $\pi^*$  ( $q \leq p$ ). If  $q$  is null, the sampling procedure is completed; otherwise a second step, referred to as the "landing phase", needs to be initiated. It consists in relaxing the balancing constraints as little as possible, and re-initialising a flight phase with these new constraints until a sample is obtained. It is not possible to change the space of the constraints from the outset in a way that might mix the vertices of  $K$  with  $C$ , as this would amount testing all possible samples to first see whether one of them allows the constraints to be met. Changing the constrained space in a later phase (the landing phase) makes it possible to work on a population  $U^*$  of smaller size ( $\dim(U^*) = q$ ). The problem can thus be solved because the number of samples to be considered is reasonable.

The implementation of this algorithm is available in SAS thanks to macro *FAST CUBE* or in package *BalancedSampling* in R.

### The local cube method

The general idea of the spatially balanced sampling algorithm is to build a cluster of  $p + 1$  geographically close units, and then to apply the cube flight phase to this cluster. This leads to decide whether a unit is selected in this cluster or not, while respecting  $p$  local constraints within

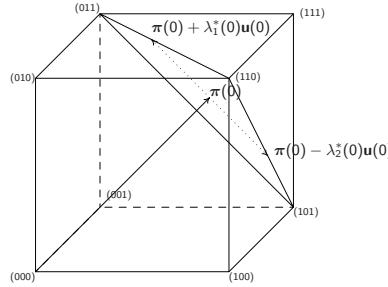


Figure 10.4 – First step of the cube flight phase for  $N = 3$  and a constraint ( $p = 1$ ) of a fixed sample size  $n = 2$

the cluster. Next, the probabilities are modified locally, ensuring that the inclusion probabilities of the nearby units are reduced if the unit on which a decision has been made is selected. This lowers the probability that one of its nearby units is selected in the following step of the algorithm. Then, the procedure is repeated: a unit is selected, followed by a cluster of  $p + 1$  units around it, and we apply the cube flight phase with the inclusion probabilities updated in the previous step. The process is repeated as long as there are still more than  $p + 1$  units. Finally, the traditional cube landing phase is applied.

The spatially-balanced sampling method described above is available in package *BalancedSampling* in R. This package, developed in C++, allows the algorithm to be applied very quickly.

### Balancing on moments

The definition of a spatially balanced sample suggests a different use of the cube algorithm for spatial sampling. For Marker et al. 2009, "a sample is spatially balanced if the spatial moments of the localised samples match the spatial moments of the population. Spatial moments are the centre of gravity and inertia." In the cube algorithm terminology, this definition may result in selecting a balanced sample on variables defined from the geographic coordinates:  $x_i, y_i, x_i^2, y_i^2, x_i y_i$ . In order to respect the first and second non-central moments:

$$\begin{aligned} \text{--- } T_x &= \sum_{i \in U} x_i, \\ \text{--- } T_y &= \sum_{i \in U} y_i, \\ \text{--- } T_{x^2} &= \sum_{i \in U} x_i^2, \\ \text{--- } T_{y^2} &= \sum_{i \in U} y_i^2, \\ \text{--- } T_{xy} &= \sum_{i \in U} x_i y_i. \end{aligned}$$

#### 10.3.4 Ordered spatial sampling methods

The methods in this family (Stevens Jr et al. 2004, Dickson et al. 2016, Lister et al. 2009) firstly rely on the creation of a path going through all statistical units. This path can be GRTS (*Generalized Random Tessellation Stratified*), Traveling-Salesman Problem (TSP) or a Peano curve. Given the order defined by this path, the aim is then to select a sample according to a method that excludes two nearby units, for example systematic sampling.

Other path-building methods exist (Hamilton paths, or curves filling space: Hilbert, Lebesgue). Similarly, other selection methods exclude nearby units with a given ordering, such as determinantal sampling designs (Loonis et al. 2017). The question of paths has been addressed in Chapter 2: "Codifying the neighbourhood structure". We describe here the repulsiveness properties of the systematic and determinantal sampling designs.

### The systematic sampling method

Systematic sampling is a sampling method that is simple to implement and makes it possible to carry out sampling with unequal probabilities while respecting those inclusion probabilities. This method was proposed by Madow 1949, then extended by Connor 1966, Brewer 1963, Pinciaro 1978, and Hidiroglou et al. 1980. It is very often used in practice for telephone surveys, for sampling on continuous data flows, or in sampling housing units for INSEE household surveys.

To draw a fixed size sample  $n$  respecting the inclusion probability vector  $\pi$ , we start by defining the cumulative sum of the inclusion probabilities by  $V_i = \sum_{l=1}^i \pi_l, i \in U$ , with  $V_0 = 0$ . For a fixed size sample, the result is  $V_N = n$ . The systematic sampling algorithm shown below is then used to decide on the units to be sampled.

#### Systematic sampling algorithm:

- Generate a random variable  $u$  uniformly distributed on the interval [0.1].
- For  $i = 1, \dots, N$ ,

$$I_i = \begin{cases} 1 & \text{if there is an integer } j \text{ so that } V_{i-1} \leq u + j - 1 < V_i, \\ 0 & \text{otherwise.} \end{cases}$$

Table 10.2 provides an example of the method for  $n = 3$  and  $N = 10$ .

$i$	1	2	3	4	5	6	7	8	9	10
$\pi_i$	0.2	0.2	0.3	0.3	0.4	0.4	0.3	0.3	0.3	0.3
$V_i$	0.2	0.4	0.7	1	1.4	1.8	2.1	2.4	2.7	3(=n)

Table 10.2 – An example of systematic sampling

For example, if the random number generated  $u$  is equal to 0.53, then units 2, 5 and 8 will be selected because they meet the constraints:

$$V_2 \leq u < V_3 \quad V_5 \leq u + 1 < V_6 \quad V_8 \leq u + 2 < V_9.$$

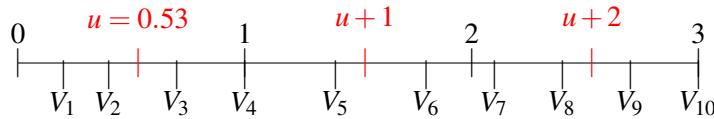


Figure 10.5 – 3 units selected out of 10

According to this method, units  $(i, j)$  respecting  $|V_i - V_j| < 1$  have zero probability of being selected together. If the file is wisely sorted, this ensures the geographical spread of the sample.

#### Implementation of the GRTS method

The GRTS method is one of the most frequently-used methods in practice when it comes to systematic sampling on a geographically-ordered file. GRTS ordering is described in chapter 2 "Codifying geographical structure".

The R *gstat* package of software R has been implemented specifically to produce samples using this method. However, the GRTS method has some drawbacks, in particular the fact that the cutting algorithm and the sampling algorithm are not dissociated, nor is the method's computational time. Indeed, the method proposes by default to stop at 11 hierarchical levels in the decomposition process, as the time needed to execute the method may be too long if a more detailed cutting is

requested. This makes it difficult for the GRTS algorithm to adapt to large populations. In order to overcome these computational limits, a new pivotal method using another tessellation algorithm (Chauvet et al. 2017) has been developed in R. In this method, the tessellation algorithm (very similar to the GRTS one) is dissociated from the sampling algorithm. This method is based on binary decomposition, making it possible to carry out the decomposition directly on 31 levels. The computational time is therefore considerably improved. In addition, it is possible to use this method in more than two dimensions.

### Determinantal sampling design

By definition, given a random variable  $\mathbb{S}$  with values in  $2^U$ , the probability law will be a determinantal sampling design if and only if there is a contracting hermitian matrix  ${}^5K$  indexed by  $U$ , referred to as a kernel, as for all  $s \in 2^U$ ,

$$p(s \subseteq \mathbb{S}) = \det(K|_s) \quad (10.5)$$

where  $K|_s$  is under the matrix of  $K$  indicated by units of  $s$ . This definition directly gives rise to the calculation of inclusion probabilities (table 10.3).

$\pi_i = pr(i \in \mathbb{S}) = \det(K _{\{i\}}) = K_{ii}$
$\pi_{ij} = pr(i, j \in \mathbb{S}) = \det \begin{pmatrix} K_{ii} & K_{ij} \\ \bar{K}_{ij} & K_{jj} \end{pmatrix} = K_{ii}K_{jj} -  K_{ij} ^2$

Table 10.3 – Calculation of simple and joint inclusion probabilities in a determinantal sampling design DSD( $K$ ) ( $|z|$  refers to the complex number module  $z$ )

The diagonal entries of  $K$  are the simple inclusion probabilities. Another particularly important result of determinantal sampling designs is the following: a determinantal design is of a fixed size if and only if  $K$  is a projection matrix<sup>5</sup>(Hough et al. 2006).

Let us consider all projection matrices in which the diagonal is a vector  $\Pi$  of inclusion probabilities *a priori*. Among them, matrix  $K^\Pi$  (whose coefficients are provided in table 10.4) offers interesting properties in terms of spatial repulsion.

The repulsiveness of the determinantal sampling design associated to  $K^\Pi$  for close statistical units (given the order listed in the file) is illustrated by the following properties (Loonis et al. 2017):

1. the design will select at most one unit within a range of the form  $]i_r + 1, i_{r+1} - 1[$ ;
2. if a unit is drawn in that range of indices, as well as the "close" unit  $i_{r+1}$ , then the design will not select an additional "close" unit, *i.e.* in  $]i_{r+1} + 1, i_{r+2} - 1[$ ;
3. this design will always have at least one individual in an interval  $[i_r + 1, i_{r+1} - 1]$ ;
4. if  $|i - j|$  is large enough, then  $\pi_{ij} \approx \Pi_i \Pi_j$  that is the joint inclusion probabilities of the Poisson design.

Applying the results of the determinantal sampling designs to the probabilities defined in table 10.2 results in quantities:  $i_1 = 4, i_2 = 7, i_3 = 10$  and  $\alpha_4 = 0.3 = \Pi_4, \alpha_7 = 0.2, \alpha_{10} = 0.3 = \Pi_{10}$ .

5. A complex matrix  $K$  is hermitian if  $K = \bar{K}^t$ , where the coefficients of  $\bar{K}$  are conjugates of those of  $K$ . A matrix is a contracting matrix if all its own values are between 0 and 1.

6. A hermitian matrix is projection if its own values are 0 or 1.

Values of $i$	Values of $j$	
	$j = i_r$	$i_r < j < i_{r+1}$
$i_{r'} < i < i_{r'+1}$	$-\sqrt{\Pi_i} \sqrt{\frac{(1-\Pi_j)(\Pi_j-\alpha_j)}{1-(\Pi_j-\alpha_j)}} \gamma_r^{r'}$	$\sqrt{\Pi_i \Pi_j} \gamma_r^{r'}$
$i = i_{r'+1}$	$-\sqrt{\frac{(1-\Pi_i)\alpha_i}{1-\alpha_i}} \sqrt{\frac{(1-\Pi_j)(\Pi_j-\alpha_j)}{1-(\Pi_j-\alpha_j)}} \gamma_r^{r'}$	$\sqrt{\frac{(1-\Pi_i)\alpha_i}{1-\alpha_i}} \sqrt{\Pi_j} \gamma_r^{r'}$

where for every  $r$  such as  $1 \leq r \leq n$ :

- $1 < i_r \leq N$  is a integer such that  $\sum_{i=1}^{i_r-1} \Pi_i < r$  et  $\sum_{i=1}^{i_r} \Pi_i \geq r$ ; by convention, it is established that  $i_0 = 0$
- $\alpha_{i_r} = r - \sum_{i=1}^{i_r-1} \Pi_i$ . It should be noted that  $\alpha_{i_r} = \Pi_{i_r}$  if  $\sum_{i=1}^{i_r} \Pi_i = r$ .
- $\gamma_r^{r'} = \sqrt{\prod_{k=r+1}^{r'} \frac{(\Pi_{i_k}-\alpha_{i_k})\alpha_{i_k}}{(1-\alpha_{i_k})(1-(\Pi_{i_k}-\alpha_{i_k}))}}$  for  $r < r'$ ,  $\gamma_r^{r'} = 1$  otherwise.

Table 10.4 – Values of  $K_{ij}^{\Pi}$  with  $i > j$

The joint inclusion probabilities are given in the matrix below.

$$\left( \begin{array}{cccccccccc} 0 & 0 & 0 & \frac{2}{25} & \frac{2}{25} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} \\ 0 & 0 & 0 & \frac{2}{25} & \frac{2}{25} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} \\ 0 & 0 & 0 & \frac{3}{25} & \frac{3}{25} & \frac{9}{100} & \frac{100}{9} & \frac{100}{9} & \frac{100}{9} \\ 0 & 0 & 0 & \frac{3}{25} & \frac{3}{25} & \frac{100}{9} & \frac{100}{9} & \frac{100}{9} & \frac{100}{9} \\ \frac{2}{25} & \frac{2}{25} & \frac{3}{25} & \frac{3}{25} & 0 & \frac{1}{20} & \frac{7}{60} & \frac{7}{60} & \frac{7}{60} \\ \frac{2}{25} & \frac{2}{25} & \frac{3}{25} & \frac{3}{25} & 0 & \frac{1}{20} & \frac{7}{60} & \frac{7}{60} & \frac{7}{60} \\ \frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{1}{20} & \frac{7}{60} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\ \frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{1}{15} & 0 & 0 & 0 \\ \frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{1}{15} & 0 & 0 & 0 \\ \frac{3}{50} & \frac{3}{50} & \frac{100}{9} & \frac{100}{9} & \frac{7}{60} & \frac{1}{15} & 0 & 0 & 0 \end{array} \right).$$

The entries around the main diagonal tend to be null or close to 0, reflecting repulsiveness.

## 10.4 Comparing methods

Various sampling methods aimed at taking spatial information into account have been presented. This section compares their relative efficiency, using real data.

### 10.4.1 The principle

The 2015 tax data is geocoded for all households, making it possible to split the territory of the Provence-Alpes-Côte d'Azur region (PACA) into 1 012 primary sampling units (PSU) of approximately 2 000 primary residences. Each of these PSUs is characterised by fifteen variables of interest describing its socio-economic or demographic situation. We focus here on the statistical properties of first-stage sampling, *i.e.* a sampling of  $m$  primary units from amongst the  $M = 1012$  PSUs.

Only the geographical coordinates of the barycentres of the PSU at the time of sample selection are available. Two sets of inclusion probabilities are tested: the first with equal probabilities, the second with probabilities proportional to the number of the unemployed. Both of these sets are tested for three different sample sizes:  $m = 30, 60, 100$ .

The aim is to assess the methods presented above by comparing their performance with those of a benchmark method. This benchmark is simple random sampling (SRS) for equal probability

designs, and systematic sampling on a randomly sorted file for unequal probabilities sampling designs. Each method is assessed through two types of indicators:

### 1. Estimating the variance

For each method, the aim is to determine to what extent the variance of the total of a given variable is reduced in respect to the variance found with the *benchmark* method. This is achieved by studying a set of variables of interests with different levels of spatial autocorrelation.

For all methods apart from the determinantal sampling design, variances of totals are estimated using the Monte Carlo method, replicating 10,000 times each method for each set of inclusion probabilities and each sample size. Concerning determinantal sampling designs, variance can be computed exactly since the joint inclusion probabilities are known.

The aim is to find out whether the gain in variance is greater when the variable is spatially autocorrelated. Therefore, the 15 variables of interest are ranked according to their level of spatial autocorrelation, measured by Moran's I dilated by inclusion probabilities, since  $\frac{y_i}{\pi_i}$  determines the quality of the results according to Equation 10.1. When the design is an equal probability sampling design, it is similar to computing directly Moran's I for each variable (table 10.5).

### 2. The Voronoï indicator

For each method, an empirical dispersion indicator (known as the Voronoï index) is also computed, by following Stevens Jr et al. 2004. The principle is as follows:

- the Voronoï diagram is built only with the  $m$  selected PSUs;
- for a given selected PSU  $i$ , PSUs located in the Voronoï polygon associated with  $i$  are identified from amongst the 1 012 original PSUs;
- the sum  $\delta_i$  of these PSUs' inclusion probabilities is computed. The average of the  $\delta_i$  is equal to 1, since the sum of the inclusion probabilities over the 1 012 PSU is  $m$  and because the  $m$  polygons partition the territory. If the procedure has selected only few units around a given selected PSU  $i$ ,  $\delta_i$  will be greater than 1. If the procedure has selected a lot of other units close to  $i$ ,  $\delta_i$  will be less than 1 (see Figure 10.6);
- for a random sample  $\mathbb{S}$ , the Voronoï indicator is then defined by:

$$\Delta_{\mathbb{S}} = \frac{1}{m-1} \sum_{i \in \mathbb{S}} (\delta_i - 1)^2.$$

The more uniformly a procedure spreads the units, the lower the dispersion of  $\delta_i$  measured by  $\Delta_{\mathbb{S}}$  will be. The expected value of  $\Delta_{\mathbb{S}}$  will be estimated by simulation (average over 10 000 replications, noted  $V$ ).

The Voronoï index can be computed in R using the function `sb()` of package *BalancedSampling* or based on the R codes provided in Benedetti et al. 2015 (pp. 161-162).

#### 10.4.2 Results

Ten spatial sampling methods are studied:

- 4 methods in the so-called “A” family in the following. Family A consists in updating inclusion: Poisson sampling, local pivotal, local cube<sup>7</sup> and balanced cube on spatial moments;
- 6 methods in a second so-called “B” family of methods. Family B methods are based on prior ordering of the file. In total, 3 paths are considered (Figure 10.7): TSP (10.7a), Hamilton (10.7b), and GRTS (10.7c), and each followed by a systematic sampling or a determinantal sampling. These three paths are computed using an exact method. Then, all the sampling replications run on a single sorted file.

7. The local pivotal and local cube are equivalent methods in this context.

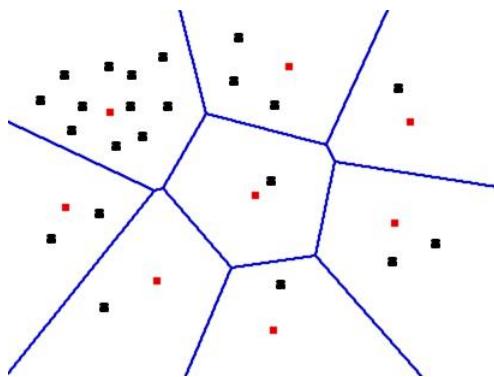


Figure 10.6 – Calculation of the Voronoï index

**Note:** Voronoï polygons are built around the selected units (red).  $\delta_i$ s are calculated on all units (red and black).

Variable	Moran I $\pi$ constant	Moran's I dilated ( $\frac{y_i}{\pi_i}$ )
Number of households earning agricultural income	0,68	0,66
Total wage income	0,62	0,55
Number of couples with child(ren)	0,61	0,54
number of those receiving minimum social benefits	0,60	0,61
Number of poor	0,58	0,58
Number of children	0,55	0,52
Number of people living in a neighbourhood targeted by City Policy	0,55	0,54
Number of households owning their homes	0,52	0,47
Total standard of living	0,46	0,46
Number of unemployed	0,45	0,42
Number of single-parent families	0,41	0,43
Number of individuals	0,40	0,34
Number of men	0,39	0,34
Number of women	0,24	0,34
Number of households	0,08	0,38

Table 10.5 – Moran's indices for different variables computed at the PSU level of PACA

**Source:** INSEE, *Fideli 2015*

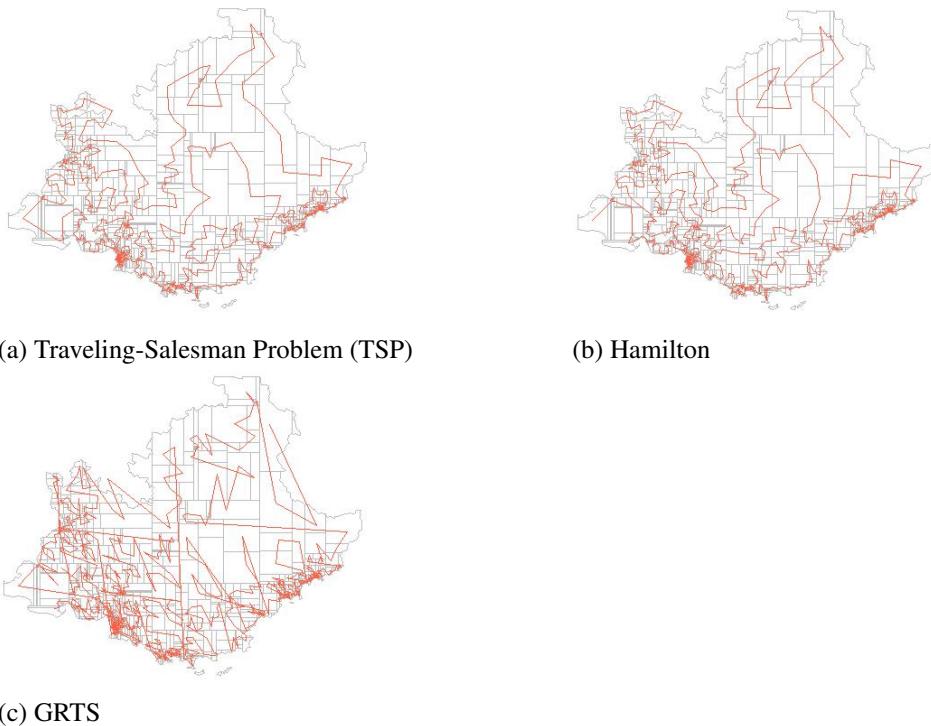
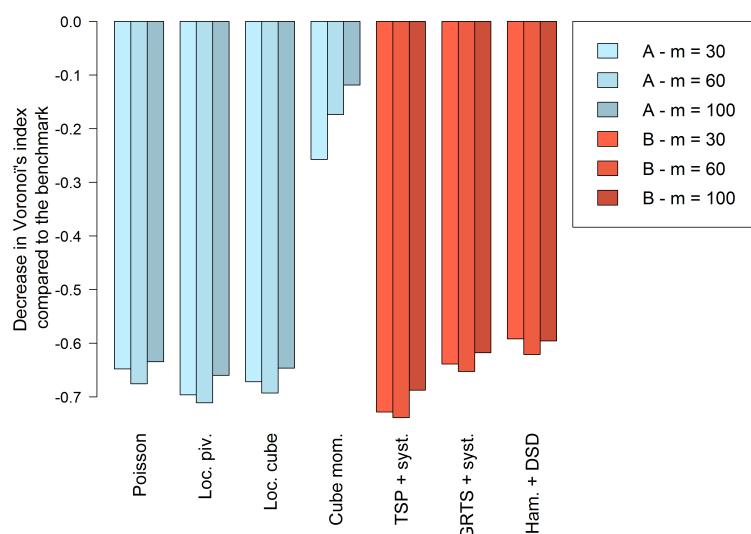


Figure 10.7 – Paths connecting PSU centroids

Source: INSEE, Fideli 2015

Figure 10.8 –  $(V^q - V^{ref})/V^{ref}$ , where  $V^q$  is the Voronoï index for method  $q$  and  $V^{ref}$  for benchmark, for different values of  $m$  (example offered by equal probabilities)

Source: INSEE, Fideli 2015

**Note:** For a sampling with 30 PSUs using the Poisson sampling method with equal probabilities, the Voronoï indicator (averaged over 10 000 replications) is 65% less than the single random sampling (benchmark).

Figure 10.8 provides  $(V^q - V^{ref})/V^{ref}$ , where  $V^q$  stands for the Voronoï index for method  $q$  and  $V^{ref}$  stands for the same indicator for benchmarking. A significantly negative value reveals a better spatial dispersion. The figure shows that for all methods and sample sizes, the Voronoï index

is significantly improved: by -60 to -70% compared to the benchmark. Only the balanced moments method is less efficient.

For a given method and sample size, Figures 10.9 and 10.11 represent, as they do for the Voronoï index, the decrease of variance of a variable of interest, in comparison to the benchmark. This decrease is related to the intensity of the spatial autocorrelation of the variable dilated by inclusion probabilities.

For the methods shown in figure 10.9, *i.e.* most of the methods studied, the gain in terms of variance is all the greater as the variable is spatially autocorrelated. However, this result is clearer with equal probabilities (10.9a) than with unequal probabilities (10.9b). These methods are equivalent in terms of gain. Consequently, Poisson sampling, local pivotal, local cube and determinantal designs on ordered file (TSP or Hamilton ordering) almost halve the variance of the sample, for the most autocorrelated variables and for  $m = 100$ . Furthermore, for all methods represented in Figure 10.9, the relative gain in variance is all the greater as the sampling rate is higher. Figure 10.10 illustrates this result for determinantal plans with equal probabilities.

The four methods shown in red and blue in Figure 10.11 lead to results different from methods presented in Figure 10.9:

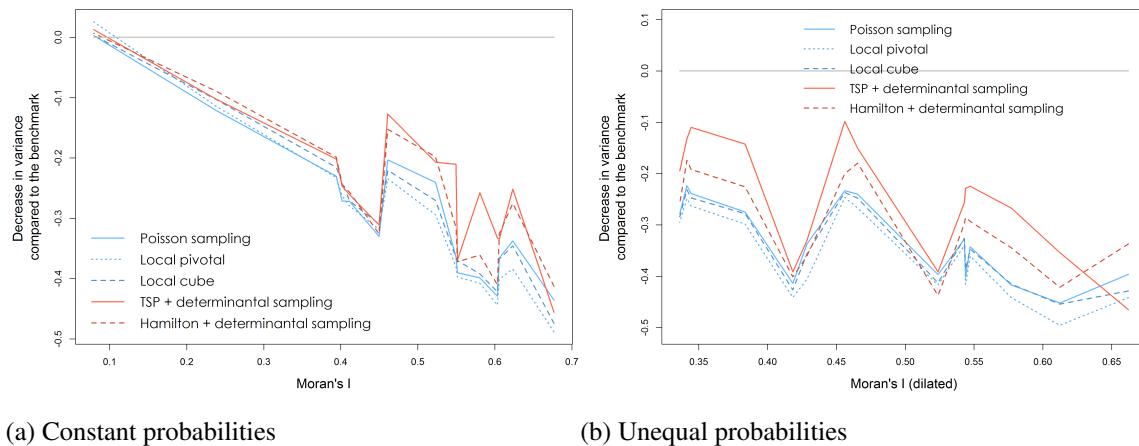
- the cube method balanced on first and second-order spatial moments ( $x, y, x * y, x^2$  and  $y^2$ , where  $x$  and  $y$  are spatial coordinates) is less effective in terms of gain in variance. Calibrating with the inertia of the total population finally reproduces in the sample the groupings and repulsions of units. That goes against the desired sample dispersion principle;
- file ordering (TSP, Hamilton or GRTS) followed by a systematic sampling, yields more erratic results than other methods in the same family. Entropy<sup>8</sup> of the systematic survey design is very weak, and this is even more the case on a uniquely sorted file. The number of potential samples with this method is  $M/m$ , explaining why curves in Figure 10.11 look less smooth and why it is more difficult to draw conclusions. However, these methods still perform very well in terms of sample dispersion. In particular, the TSP ordering followed by systematic sampling is the one that reduces the Voronoï indicator the most (Figure 10.8). It is also the one that most reduces the variance of the variables most spatially autocorrelated. GRTS ordering, meanwhile, is less efficient, due to lower ordering quality (the total length of the path obtained with GRTS is almost twice as long as the TSP or Hamilton path, see Figure (10.7)).

## Conclusion

The creation of samples from a georeferenced sampling frame offers a possible new context for judicious mobilisation of geographical information. This chapter has presented various methods using this information at different stages of the sampling design process. We have carried out some tests based on real data, aiming at comparing these methods with traditional or original precision indicators, and testing different sets of parameters. The large majority of the suggested methods prove to be effective in that they yield accurate estimates, even though the systematic sampling methods appear less effective. The statistical efficiency of a spatial sampling method increases with the level of spatial autocorrelation in the variable of interest to be estimated.

---

8. Entropy is a measure of disorder. A high-entropy design enables a large number of samples to be selected and therefore leaves significant room for randomness



(a) Constant probabilities

(b) Unequal probabilities

Figure 10.9 – Decrease in variance vs. benchmark for different methods, according to the spatial autocorrelation index of the variable (example with  $m = 60$ )

**Source:** INSEE, *Fideli* 2015

**Note:** Each curve stands for a spatial sampling method, and each point of the curve reflects 10 000 samples taken using the same method. The variation in variance of a given variable relative to a benchmark (in percentage) is shown, depending on the variable's level of spatial autocorrelation. For example, for an equal probabilities sample of 60 PSUs with the Poisson sampling method, the variance of the "number of women" variable (Moran's I = 0.24) is 11% less than with SRS.

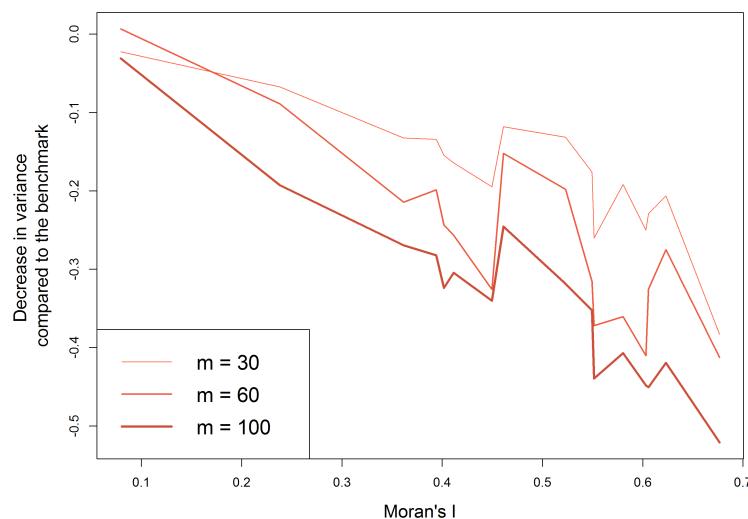


Figure 10.10 – Reductions in variance vs. benchmark according to the spatial autocorrelation index of the variable, for different values of  $m$  (e.g. determinantal sampling with Hamilton ordering processes and equal probabilities)

**Source:** INSEE, *Fideli* 2015

**Note:** When using determinantal sampling with equal probabilities, the variance of the variable "number of women" (Moran's I = 0.24) is decreased by 6.7% for a sample of 30 PSUs, by 8.9% for a sample of 60 PSUs, and by 19.3% for a sample of 100 PSUs, compared to simple random sampling.

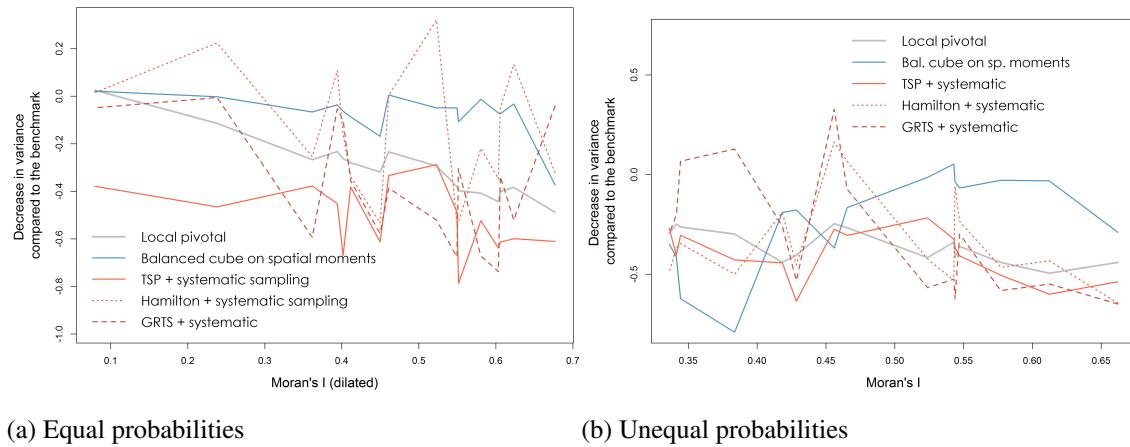


Figure 10.11 – Reductions in variance vs. *benchmark* for different methods, according to the spatial autocorrelation index of the variable (example with  $m = 60$ )

**Source:** INSEE, *Fideli* 2015

**Note:** The figure's local pivotal method 10.9 is represented in a grey line for comparison purposes.

## References - Chapter 10

- Ardilly, Pascal (2006). *Les techniques de sondage*. Editions Technip.
- Benedetti, Roberto, Federica Piersimoni, and Paolo Postiglione (2015). *Sampling Spatial Units for Agricultural Surveys*. Springer.
- Bondesson, Lennart and Daniel Thorburn (2008). « A List Sequential Sampling Method Suitable for Real-Time Sampling ». *Scandinavian Journal of Statistics* 35.3, pp. 466–483.
- Brewer, K.R.W. (1963). « A model of systematic sampling with unequal probabilities ». *Australian & New Zealand Journal of Statistics* 5.1, pp. 5–13.
- Chauvet, Guillaume and Ronan Le Gleut (2017). « Asymptotic results for pivotal sampling with application to spatial sampling ». *Work in progress*.
- Connor, W.S. (1966). « An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement ». *Journal of the American Statistical Association* 61.314, pp. 384–390.
- Deville, Jean-Claude and Yves Tillé (1998). « Unequal probability sampling without replacement through a splitting method ». *Biometrika* 85.1, pp. 89–101.
- (2004). « Efficient balanced sampling: the cube method ». *Biometrika* 91.4, pp. 893–912.
- Dickson, Maria Michela and Yves Tillé (2016). « Ordered spatial sampling by means of the traveling salesman problem ». *Computational Statistics*, pp. 1–14. DOI: 10.1007/s00180-015-0635-1. URL: <http://dx.doi.org/10.1007/s00180-015-0635-1>.
- Favre-Martinoz, Cyril and Thomas Merly-Alpa (2017). « Constitution et Tirage d'Unités Primaires pour des sondages en mobilisant de l'information spatiale ». *49<sup>ème</sup> Journées de statistique de la Société Française De Statistique*.
- Ganganath, Nuwan, Chi-Tsun Cheng, and K Tse Chi (2014). « Data clustering with cluster size constraints using a modified k-means algorithm ». *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on*. IEEE, pp. 158–161.
- Grafström, Anton (2012). « Spatially correlated Poisson sampling ». *Journal of Statistical Planning and Inference* 142.1, pp. 139–147.
- Grafström, Anton and J Lisic (2016). « BalancedSampling: Balanced and spatially balanced sampling ». *R package version 1.2*.

- Grafström, Anton, Niklas LP Lundström, and Lina Schelin (2012). « Spatially balanced sampling through the pivotal method ». *Biometrics* 68.2, pp. 514–520.
- Grafström, Anton and Yves Tillé (2013). « Doubly balanced spatial sampling with spreading and restitution of auxiliary totals ». *Environmetrics* 24.2, pp. 120–131.
- Hidiroglou, M.A. and G.B. Gray (1980). « Algorithm AS 146: Construction of Joint Probability of Selection for Systematic PPS Sampling ». *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29.1, pp. 107–112.
- Hough, J Ben et al. (2006). « Determinantal processes and independence ». *Probab. Surv* 3, pp. 206–229.
- Lisic, Jonathan (2015). « Parcel level agricultural land cover prediction ». PhD thesis. George Mason University.
- Lister, Andrew J and Charles T Scott (2009). « Use of space-filling curves to select sample locations in natural resource monitoring studies ». *Environmental monitoring and assessment* 149.1, pp. 71–80.
- Loonis, Vincent (2009). « La construction du nouvel échantillon de l’Enquête Emploi en Continu à partir des fichiers de la Taxe d’Habitation. » *JMS* 2009, p. 23.
- Loonis, Vincent and Xavier Mary (2017). « Determinantal sampling designs ». *arXiv preprint arXiv:1510.06618*.
- Madow, William G (1949). « On the theory of systematic sampling, II ». *The Annals of Mathematical Statistics*, pp. 333–354.
- Malinen, Mikko I and Pasi Fränti (2014). « Balanced k-means for clustering ». *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, pp. 32–41.
- Marker, David A and Don L Stevens (2009). « Sampling and inference in environmental surveys ». *Handbook of Statistics* 29, pp. 487–512.
- Pinciaro, Susan J (1978). « An algorithm for calculating joint inclusion probabilities under PPS systematic sampling ». *of: ASA Proceedings of the Section on Survey Research Methods*, pp. 740–740.
- Stevens Jr, Don L and Anthony R Olsen (2004). « Spatially balanced sampling of natural resources ». *Journal of the American Statistical Association* 99.465, pp. 262–278.
- Tai, Chen-Ling and Chen-Shu Wang (2017). « Balanced k-Means ». *Asian Conference on Intelligent Information and Database Systems*. Springer, pp. 75–82.



# 11. Spatial econometrics on survey data

RAPHAËL LARDEUX, THOMAS MERLY-ALPA  
INSEE

---

<b>11.1 First approach by simulation</b>	<b>280</b>
11.1.1 Simulation of a SAR . . . . .	280
11.1.2 Sampling procedures . . . . .	282
11.1.3 Results and interpretation . . . . .	284
11.1.4 A "size effect" . . . . .	285
11.1.5 Robustness . . . . .	287
<b>11.2 Prospects for resolution</b>	<b>288</b>
11.2.1 Moving to a higher scale by aggregation . . . . .	288
11.2.2 Imputing missing data . . . . .	289
<b>11.3 Empirical application: the manufacturing sector in Bouches-du-Rhône</b>	<b>291</b>
11.3.1 Data . . . . .	292
11.3.2 Identification . . . . .	292
11.3.3 Estimation . . . . .	293
11.3.4 Spatial estimates on samples . . . . .	294
11.3.5 Estimation on aggregated data . . . . .	296
11.3.6 Imputation of missing data . . . . .	297

---

## Abstract

Spatial econometrics requires comprehensive data on a territory, which in principle prohibits the use of survey data. This chapter presents the pitfalls in estimating spatial autoregressive models (SAR) on sampled data and assesses potential corrections offered by the empirical literature. We identify two sources of bias: (*i*) a "size effect" resulting from the distortion of the spatial weighting matrix and (*ii*) a "sparsity effect" resulting from the omission of units spatially correlated with the units observed. Both effects tend to underestimate the magnitude of spatial correlations. However, the bias is lower in the case of a cluster survey and when the sample is large enough. Two types of methods are offered by the empirical literature to sidestep these pitfalls: imputing missing values (linear regression, hot deck) and aggregating data on a higher scale. These potential solutions are efficient under very restrictive conditions, part of the difficulty being the reconstruction of complex information based on few observations. The last part of the chapter illustrates this issue with the estimation of production externalities between plants of the manufacturing sector located in the French department of Bouches-du-Rhône.

- R** Prior reading of Chapters 2: "Codifying the neighbourhood structure", 3: "Spatial autocorrelation indices" and 6 "Spatial econometrics: common models" is recommended.

## Introduction

Recent developments in spatial econometrics and geolocation have made it possible to analyse spatial phenomena at very local scales. Concepts derived from spatial econometrics analysis are used in an increasingly diverse range of fields such as geostatistics, economics and network analysis. However, the application of these spatial analysis methods requires comprehensive data, which is not always accessible (due in particular to non-response, excessively lengthy collection time, ...) and cannot easily be processed within the limited time available. The extension of spatial econometrics to survey data would make it possible to take full advantage of a detailed piece of information to produce a fine-grained metric of the incidence of spatial correlations on econometric estimates<sup>1</sup>.

In this chapter, we discuss recent developments in the application of spatial estimation methods when a part of the data set is missing, particularly in the case of survey data. We show that estimating spatial econometric models on survey data underestimates spatial correlations, in particular when data is collected according to simple random sampling and when the sample is small. Ignoring missing observations is never an appropriate solution, but other potential correction methods such as imputing missing data or switching to a coarser scale through data aggregation may be efficient under very restrictive assumptions. We will not be addressing the possibility of a spatial survey, which is particularly complex in the case of social data<sup>2</sup>. Neither will we develop the case of unknown location.

Why does spatial econometrics require comprehensive data? Traditional econometrics is based on the hypothesis of mutual independence between observations. Estimating a model on a data subset may affect the power of statistical testing but, in the absence of a selection problem, estimators remain unbiased and efficient. In contrast, in spatial econometrics models, observations are considered to be correlated with each other: each unit is influenced by its neighbours. Removing some observations amounts to omitting their links with nearby units, which biases the spatial correlation parameter as well as spatial effects. We show that this bias minimises the value of the spatial correlation parameter, since some neighbourhood links are no longer taken into account in the estimate.

Conceptually, spatial econometrics differs from traditional econometrics because of the assumptions about the data generating process. In traditional econometrics, observations are considered as a representative random sample of a population and are interchangeable. Spatial analysis conceives of them as the sole completion of a spatial process, each observation being necessary to estimate the underlying process<sup>3</sup>. Spatial econometrics has been developed within the framework of Cliff et al. 1972, characterised by exhaustive and perfect information on spatial units and by the absence of missing data (Arbia et al. 2016). In practice, these conditions are almost never met. Therefore, direct application of spatial estimation techniques to sampled data can significantly affect the results.

The estimation of spatial models on survey data induces several issues. First, estimates are disrupted by a "size effect". The existence of  $m$  missing data in a population of size  $n$  gives rise to a weighting matrix of size  $(n - m) \times (n - m)$  instead of the effective weighting matrix of size  $n \times n$ . This dimension reduction skews in turn the estimated spatial correlation parameter (Arbia et al. 2016). Second, estimation on an incomplete data set implies that interconnections

1. Pinkse et al. 2010 refer to this prospect as "the future of spatial econometrics".

2. On these questions, readers are invited to refer to Chapter 10 "Spatial sampling".

3. In this sense, spatial analysis is similar to time series, where the observed dataset is derived from a stochastic process.

between observed and missing units will be overlooked, which induces a measurement error on the neighbourhood (regressor  $WY$ ) and biases estimated parameters. We compare the consequences of both effects estimating SAR models on different samples drawn from a population which has been simulated according to the exact same SAR specification. We show that beyond the "size effect", the "sparsity effect" has significant consequences.

While a number of corrections have been proposed, none of them are adequate to the current framework<sup>4</sup>. When the location of individuals is known, imputation is generally preferred (Rubin 1976; Little 1988; Little et al. 2002). However, naive imputation, for example through linear models, is not efficient to provide unbiased estimates (Belotti et al. 2017a). To get around this problem, Kelejian et al. 2010 develop estimators when only an incomplete subset of a population is available. Wang et al. 2013a suggest an imputation method through two-stage least squares, in a setting where the values of the dependent variable are randomly missing. In the same framework, LeSage et al. 2004 rely on the EM algorithm (Dempster et al. 1977). First, step "E" (expected) assigns a value to the missing data, conditionally on observables and parameters of the spatial model. Then step "M" (maximisation) determines the value of these parameters by likelihood maximisation. By iteration, this procedure makes it possible to draw from an estimated model all the information available to impute missing values. More recent work by Boehmke et al. 2015 extends this procedure to the case of missing observations (unknown dependent and independent variables).

Recent empirical papers illustrate the importance of these corrections. In a hedonic price model, LeSage et al. 2004 rely on the EM algorithm to predict the value of unsold housing. In a network model with space autocorrelation, Liu et al. 2017 show that the detection of a peer effect requires taking the sampling process into account. Yet, such complex imputation methods based on an estimated model (*model-based*) are still rarely applied. When some data is missing, the solution generally chosen is to remove the corresponding observations from the field of the estimation. The risk is that an attenuation bias in the spatial correlation may be generated. Some estimations are limited to a subset, in particular, a specific region or group, leading to a potential "size effect" as well as an underestimation of correlations at the edge of the considered area (Kelejian et al. 2010). Lastly, while most applications are performed on aggregated data to benefit from comprehensive data on a larger scale, this solution can induce positional errors<sup>5</sup> (Arbia et al. 2016) as well as an ecological bias (Anselin 2002b). We discuss the impact of these various methods on spatial estimates.

The issue of missing values in a framework where observations are correlated has been highlighted by other fields of statistics related to spatial econometrics: time series, geostatistics and network econometrics. Time series and geostatistics are similar to continuous spatial data processing. The issue of missing data was addressed very early in the field of time series (Chow et al. 1976, Ferreiro 1987). Jones 1980; Harvey et al. 1984 recommend using a Kalman filter to concurrently estimate a model and impute values. Geostatistic analysis corrects incomplete data sets either upstream using spatial sampling methods, or by predicting the value of a continuous spatial variable in an unknown position (spatial interpolation or kriging, see Chapter 5: "Geostatistics"). Longitudinal approaches combining kriging with the Kalman filter have also been developed (Mardia et al. 1998).

4. In particular, these methods vary according to the underlying assumptions on missing data. Depending on the value and/or location of the observations, the dependent and/or independent variables are affected and depending on whether the likelihood of a piece of data going missing depends on the correlations with observable and/or unobservable data. The literature on the incidence of missing data thus establishes a distinction between *Missing at Random* (MAR), *Missing Completely at Random* (MCAR) and *Missing Not at random* (MNAR). cf Rubin 1976, Huisman 2014

5. Arbia et al. 2016 develop this concept to refer to cases where the position of an observation ( $X, Y$ ) is not known precisely. For example, lack of precision in measurement, metric blurred for confidentiality reasons, missing addresses.

However, continuous data methods cannot be transposed straightforwardly into economic and social analysis, where data is fundamentally discrete. Furthermore, the use of such spatial survey techniques would contradict fundamental principles of social data collection, such as equi-weighting and the use of deterministic sampling bases. Network econometrics focuses on the issue of missing observations (Burt 1987; Stork et al. 1992; Kossinets 2006). Estimation of spatial autocorrelation on a sample of a network is gaining momentum with the growing use of social networks (Zhou et al. 2017). However, practical solutions remain rare. As in spatial econometrics, the main difficulty is to reconstruct information on unobserved units based on observed data, without knowing the effect of the former on the latter (Koskinen et al. 2010). In particular, Huisman 2014 does not come out clearly in favour of any traditional imputation strategy and remains very cautious about extensions of imputation methods to network data. Solutions based on sampling methods have also been proposed in order to collect data on populations of interest (Gile et al. 2010).

This chapter focuses on two questions: which biases are generated by the estimation of spatial econometric models on survey data? What are the consequences of classic solutions aimed at correcting missing data (data deletion, imputation, aggregation)? These questions have been addressed by Arbia et al. 2016, who proceed by simulation and observe a stronger incidence of missing data when these are grouped in clusters, in which case all local phenomena may be lost. However, they consider cases where missing data accounts for at most 25% of the population, which is very low compared to survey data, where they generally reach more than 90% of the population.

Section 11.1 highlights the bias resulting from the application of spatial methods to a non-exhaustive sample and discusses its magnitude depending on the percentage of observations sampled and the sampling method. Section 11.2 shows the consequences of some usual solutions: shifting to a higher level by aggregation and the imputation of missing values. Section 11.3 illustrates these biases through the estimation of production externalities between industries of the French department of Bouches-du-Rhône.

## 11.1 First approach by simulation

In this section, we show that the estimation of a spatial autoregressive model (SAR) on sample data is biased. In order to do that, we proceed through Monte Carlo simulations. First, we simulate a spatial data set across a geographic area so that units are correlated according to a given value of the spatial correlation parameter. Second, we draw samples from this data set and estimate the value of the spatial correlation parameter for each one of them.

### 11.1.1 Simulation of a SAR

The geographical space chosen is a map of Europe<sup>6</sup>, detailed at the administrative level NUTS3 (the lowest level in the NUTS hierarchy defined by Eurostat, which corresponds to small areas where specific studies can be carried out, such as the French departments) from which the furthest islands and Iceland have been removed in order to maintain a homogeneous and compact geographical space. From the *shapefile* of Europe, we build a neighbourhood matrix  $\mathbf{W}$  based on distance, so that the weight associated with two neighbouring units decreases according to the square of the distance and is cancelled when this distance exceeds a limit threshold. Residuals and an explanatory variable are drawn from Gaussians:  $\varepsilon \sim \mathcal{N}(0, 1)$  and  $X \sim \mathcal{N}(5, 2)$ , making it possible to ultimately simulate a variable  $Y$  following a SAR model (*Spatial Auto-Regressive*):

$$Y = (1 - \rho \mathbf{W})^{-1} X \beta + (1 - \rho \mathbf{W})^{-1} \varepsilon \quad (11.1)$$

6. This card is shared on the site: <http://ec.europa.eu/eurostat/fr/web/gisco/geodata/reference-data/administrative-units-statistical-units>.

with  $\beta = 1$  and  $\rho = 0.5$ , reference parameters which we try to find by estimating the exact same SAR model on samples. Data from simulated variables  $Y$  is shown in Figure 11.1. The presence of concentrated coloured areas is characteristic of the positive spatial autocorrelation resulting from the data-generating process.

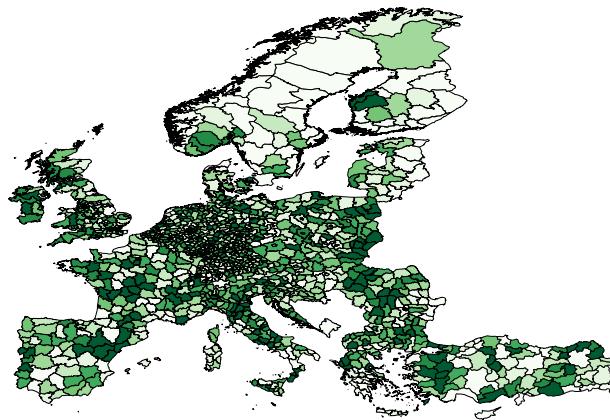


Figure 11.1 –  $Y$  simulated using a SAR model

**Copyright:** EuroGeographics for administrative boundaries

Table 11.1 shows the results of the estimation of a SAR model across all NUTS3 zones in Europe. They confirm the validity of this simulation, since the estimated parameters  $\beta$  and  $\rho$  are very close to the values initially calibrated.

$\beta$	$\rho$	Direct	Indirect	Total
0.989	0.494	1.043	0.860	1.902

Table 11.1 – Parameters estimated by SAR across all zones

**Box 11.1.1 — Simulation of a SAR with with R.** To simulate a SAR in R, the most important step is formatting its neighbourhood matrix  $\mathbf{W}$ :

```
D <- nb2listw(W, style="W", zero.policy=TRUE)
```

Once the neighbourhood matrix is in the format `listw`,  $1 - \rho \mathbf{W}$  must be inverted using the following function, of which  $\rho$  is one of the parameters:

```
InvD <- invIrW(D,rho)
```

Note: this step may be time-consuming. Then, all we need to do is simulate our variable  $Y$ :

```
Y <- (InvD %*% X) + (InvD %*% eps)
```

### 11.1.2 Sampling procedures

The challenge lies in examining the capacity of spatial models to correctly estimate  $\rho$  and  $\beta$  on samples drawn from this simulated data. In particular, we discuss the effect that sampling some of these areas may have on the estimation of the underlying model.

A survey consist in randomly selecting, using a procedure referred to as a sampling plan, a set of  $n$  units within a population of  $N$ , where  $n$  is often much smaller than  $N$  in order to limit the costs of collecting information. Survey theory states that estimates made using the sample extend without bias to the total population, but are more precise when the sample size increases and when the sampling plan is suited to the estimated variable. To explore questions around surveys, it is advised that readers refer to Ardilly 1994, Tillé 2001 or Cochran 2007.

In the rest of this section, we present a number of conventional sampling techniques and how they can be applied within the framework of the European NUTS3. However, we can already make a number of general hypotheses and comments, following the ideas developed in Goulard et al. 2013 regarding the new French population census. On the one hand, the effect should obviously not be the same according to size  $n$  of the selected sample. With just under a dozen zones, the initial spatial structure will not be able to be reconstructed, while sampling 95% or even 99% of the zones should make it easily recoverable. On the other hand, the question of the sampling method will also need to be addressed. Is the spatial dimension taken into account in the method? We can refer to Chapter 10 "Spatial sampling" to delve deeper into these issues.

#### Simple random sampling

Simple random sampling consists in drawing independently and without replacement a number  $n$  of marbles from a large bowl  $N$ . Under such conditions, all individuals have the same chance of being selected in the sample. Where one individual has been selected in a sample, this reduces the likelihood that others will also be included. In our case,  $n$  areas are selected in an entirely random manner. Figure 11.2 shows an example of a sample.



Figure 11.2 – A sample drawn according to simple random sampling ( $n = 500$ )

**Copyright:** EuroGeographics for administrative boundaries

#### Poisson sampling

Poisson (or Bernoullian) sampling consists in flipping a coin to determine, based on heads or tails, whether each individual should be included in the population. Under such conditions, all

individuals have the same chance of being selected in the sample. While an individual's being selected for a sample does not affect the likelihood that the others will be included, the sample size is not set beforehand. In our case, each zone has a likelihood  $p$  of being retained in the sample. The resulting sample size is then  $pN$  in expected terms.

### Cluster sampling

Cluster (or areolar) sampling consists in selecting groups of individuals together. Individuals always have the same chance of being selected in the sample. However, the selection of an individual within a sample has a strong impact on the likelihood that the others will also be included, as individuals of the same cluster are always selected together. Here, the process consists in combining NUTS3 zones into different clusters, then making random selection of some of these clusters. The main interest is to limit collection costs, at the expense of a loss in precision due to intra-cluster homogeneity.

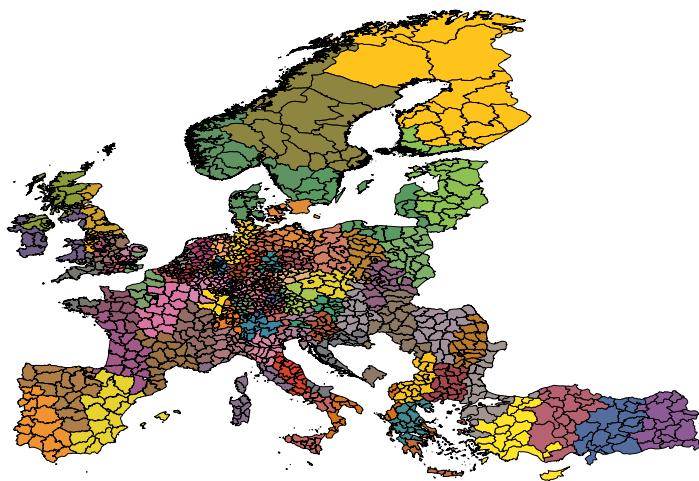


Figure 11.3 – Division of Europe into clusters

**Copyright:** EuroGeographics for administrative boundaries

It would be possible to use the different NUTS1 or NUTS2 levels as clusters. However, they are large in size and do not all include the same number of NUTS3. The problem is that large clusters limit the number of possible simulations. In contrast, clusters with different numbers of zones introduce either an issue of different sampling weights between individuals, which we do not wish to address here (see Davezies et al. 2009 for a discussion on the use of sampling weights in econometrics), or a problem of variable sample size, which may result in effects that are too complex to analyse. We therefore form clusters of the same size while maintaining a certain geographical consistency. As the weighting matrix is based on geographical distance, we give preference to the least-extensive clusters possible.

In order to obtain the same-size clusters, the number of clusters must be a divisor of the number of NUTS3 zones. In order to limit the size of the clusters, we bring together the 1,445 NUTS3 zones into 85 clusters of 17 zones each. For this purpose, we use an algorithm to build the clusters. From the area furthest away from the centre of the map, we aggregate the areas closest to it until we reach 17. As the clusters are built one by one, the most remote NUTS3 zones will already be

assigned to the construction of the previous clusters, and the algorithm will continue with more central zones. The resulting clusters are shown in Figure 11.3.

### Stratified sampling

In a stratified sampling,  $n$  units are also drawn. The difference is that  $n_1$  units are drawn from a first stratum,  $n_2$  from a second, etc.  $n_H$  in a  $H$ th, where  $n = n_1 + n_2 + \dots + n_H$ . To perform stratified sampling, it is important that the  $H$  strata be well defined, firstly, and that the allocation  $(n_1, \dots, n_H)$  be well-selected, secondly. A classical allocation is Neyman's allocation, the property of which is to minimise the variance of the estimator of the total of a variable of interest (see, for example Tillé 2001). The formula is as follows:

$$n_h = n \frac{N_h S_h}{\sum_{i=1}^H N_i S_i} \quad (11.2)$$

with  $n$  the size of the total sample,  $N_h$  the size of the stratum  $h$  and  $S_h$  the dispersion of the variable of interest within the stratum  $h$ . In some cases, when the behaviours with regard to the variable of interest are heterogeneous, this formula may give reason to run an exhaustive sampling of certain strata, *i.e.* to apply to them a 100% sampling rate.

#### 11.1.3 Results and interpretation

In order to estimate the effect of sampling part of the European NUTS3, we use the Monte Carlo method. We thus carry out 100 simulations of  $Y$  based on a SAR model, then draw 100 samples for each of them. The exact same SAR model is estimated on each sample to retrieve the estimated values of the parameters. The parameters ultimately shown in the results are the averages of  $\rho$  and  $\beta$  out of the 10,000 samples, and their standard deviations are calculated on these 10,000 values.

For each of the 10,000 draws, we keep the values of  $X$  and  $Y$  and we reconstruct a spatial weighting matrix  $\mathbf{W}_{\text{échantillon}}$  based on distance, as previously, but limited to the units included in the sample. Different sample sizes and sampling methods are considered.

### Simple random sampling

Table 11.2 shows the results using simple random sampling for sample sizes  $n$  varying from 50 to 250 zones. Significant spatial autocorrelation may be detected for a sample size above  $n = 150$ , which corresponds in the present case to a sampling rate of 1/10. Parameter  $\beta$  is estimated without bias, regardless of the size of the sample, but estimated parameter  $\hat{\rho}$  is well below its true value  $\rho = 0.5$  used for the simulation of the data set. Therefore, for small samples, the indirect effect does not significantly differ from zero and remains far lower than that observed across the entire population. The spatial autocorrelation is largely underestimated.

### Cluster sampling

Cluster sampling makes it possible to maintain a strong geographical structure, which in our case appears beneficial for detecting spatial effects, particularly for small values of  $n$ . From the clusters shown in section 11.1.2, we carry out drawings of different numbers of clusters ranging from 3 to 15 clusters, *i.e.* 51 to 255 zones. Table 11.3 shows the results obtained for values of  $n = 17p$ , the sample size composed of  $p$  clusters.

With a cluster survey, the estimate  $\hat{\rho}$  is closer to the true value of this parameter, which lies within its confidence interval. The accuracy of the estimate clearly improves when  $n$  increases, but the estimator remains biased. Thus, contrary to the case of simple random sampling, it is possible to capture spatial interactions even with a very low survey rate of around 3%. In fact, the units surveyed are highly concentrated in space and therefore highly representative of spatial correlations.

$n$	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
50	0.043 (0.043)	1.055*** (0.125)	1.056*** (0.125)	0.016 (0.017)	1.072*** (0.128)
100	0.058* (0.031)	1.050*** (0.087)	1.052*** (0.087)	0.032* (0.019)	1.083*** (0.091)
150	0.072** (0.028)	1.049*** (0.068)	1.051*** (0.068)	0.048** (0.020)	1.099*** (0.073)
250	0.101*** (0.026)	1.051*** (0.051)	1.054*** (0.052)	0.080*** (0.023)	1.135*** (0.060)

Table 11.2 – Estimation of a SAR model on samples drawn by simple random sampling

**Note:** \*\*\* denotes significance at 1%, \*\* significance at 5% and \* significance at 10%. Standard deviations are shown between brackets.  $n$ : number of observations in the sample. These estimates come from 10,000 simulations.

$n$	$p$	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
51	3	0.309* (0.237)	1.015*** (0.091)	1.051*** (0.097)	0.441* (0.262)	1.492*** (0.310)
102	6	0.348*** (0.100)	1.017*** (0.063)	1.054*** (0.066)	0.493*** (0.188)	1.546*** (0.215)
153	9	0.363*** (0.078)	1.017*** (0.052)	1.054*** (0.055)	0.516*** (0.152)	1.571*** (0.176)
255	15	0.377*** (0.058)	1.014*** (0.038)	1.052*** (0.040)	0.541*** (0.119)	1.593*** (0.136)

Table 11.3 – Estimation of a SAR model on samples drawn by cluster

**Note:** \*\*\* denotes significance at 1%, \*\* significance at 5% and \* significance at 10%. Standard deviations are shown between brackets.  $n$ : number of observations in the sample.  $p$ : number of clusters in the sample. These estimates come from 10,000 simulations.

However, if the number of units drawn is low, then the same is true for the accuracy of the spatial correlation estimate. Therefore, the indirect effect is effectively detected, even for small samples, and its value is closer to that obtained on the total population. The estimation of geographical effects therefore appears reasonable with surveys relying on cluster sampling.

Two questions remain. First, would this cluster sampling not lead to overestimating the detection of a spatially-controlled model, even if the effect is not major on the entire population? On the one hand, as there are few values  $X$  and  $Y$ , the term  $WY$  is paradoxically quite well-known, which might encourage giving priority to this approach. Second, and this will be developed in Part 11.1.4, the gap observed between the estimated  $\hat{\rho}$  and the real value used to generate the SAR can appear surprising, even though the spatial effects are clearly detected.

#### 11.1.4 A "size effect"

The results derived from simulation may come as a surprise to econometricians. Simple random sampling can be related to the super-population model used in econometrics<sup>7</sup>. Therefore, the

7. This term is connected with the difference between design-based and model-based approaches. Under a design-based approach, we assume that the population has deterministic  $Y$  values - the usual approach. Under a model-based

estimation of a population or model parameter is usually unbiased, as long as the sampling plan is correctly specified. However, spatial autocorrelation parameter  $\rho$  does not follow this "traditional law" of sampling theory<sup>8</sup>.

Notwithstanding the question of the random selection method used for the zones on which information about  $Y$  is retrieved, we will restrict this analysis to a number of zones below that of the entire population, inducing a change in the underlying spatial structure. Intuitively, a spatial effect results from interactions between all units constituting a territory. When this effect is spatially homogeneous, omitting some units implies that we neglect their contribution to the total spatial effect, which is then underestimated. We call this first component a "size effect". Moreover, available data may be more or less scattered spatially. When observed units are too sparse, they hardly account for the structure of spatial correlations. This "sparsity effect" also leads to underestimating the spatial correlation parameter.

The question of ecological bias, *i.e.* estimation errors of spatial econometric models that come from poor spatial specification, whether in terms of data granularity (resolution) or boundary problems, is similar to this issue. Thus, it is entirely possible, when you are restricted to  $n$  zones, with  $n < N$ , to never achieve as strong a spatial effect as across the entire population.

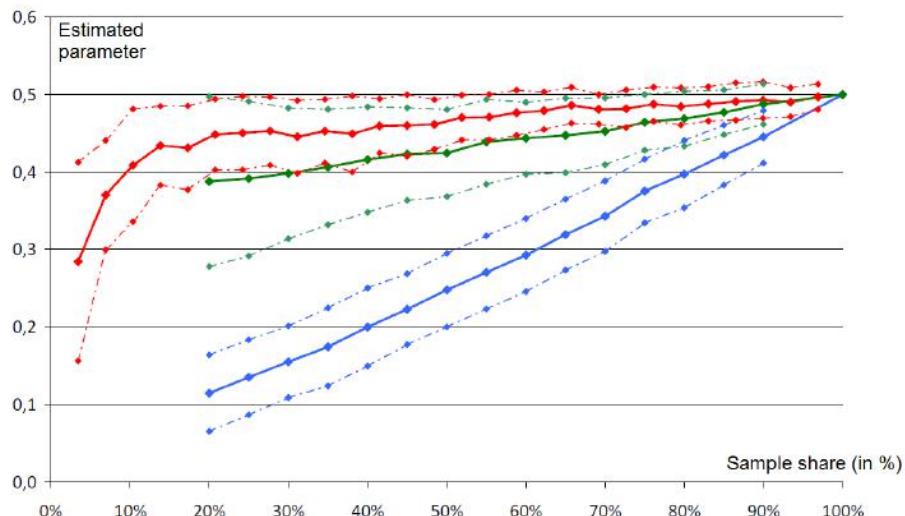


Figure 11.4 – "Size effect"

**Note:** Each point of a full-line curve represents an estimate of the parameter  $\hat{\rho}$  for a sample size stated as a percentage of the comprehensive population. When the estimate is performed on exhaustive data, we find  $\hat{\rho} = 0.5$ . The blue curve is that of a simple random survey, the green curve a cluster sampling. The red curve reflects a deterministic selection of the regions, starting from an initial point, then moving away gradually. Dotted lines represent confidence intervals at 95%.

Previous estimations were the consequence of both the size and sparsity effects. To illustrate the former, we simulate data from a SAR model on the entire population and estimate spatial correlations on a sample of the most central NUTS3  $n$ . The goal is to focus on a sub-section of

approach, we assume that there is a so-called *superpopulation* model, from which the  $Y$ s in the population are derived. Here, we are required to follow this approach in order to estimate our SAR models

8. It should be noted that multiple parameters violate this law: for example, the maximum of variable  $Y$  on a population cannot be estimated without bias from a sample. Furthermore, in our case of simple random sampling or cluster sampling, there is no problem with under-coverage, *i.e.* units of the population that cannot belong to the sample for reasons often due to the quality of the registries. This angle cannot explain the bias on  $\hat{\rho}$ .

Europe, without letting it be randomly chosen or in a fragmented manner, as was the case with previous samples (Figure 11.2). Figure 11.4 compares the values of  $\hat{\rho}$  resulting from three protocols for different percentages  $P\%$  of the total population: the selection of the most central  $P\%$  NUTS3; a cluster sampling in which each cluster of zones has  $P\%$  chances of being selected; and a classic Poisson sampling, where each area independently has  $P\%$  chances of being selected.

Just as in part 11.1.3, the Poisson sampling (similar to simple random sampling) yields estimated values  $\hat{\rho}$  much lower than the cluster survey. The main contribution of this figure is in the red curve, which is based on a non-random selection of part of the zones. It converges quicker than the others toward 0.5, the true value of  $\rho$ . This appears to confirm the hypothesis of a bias linked to the distortion of the spatial structure or "size effect", resulting from a restriction to a subset of the total population.

### 11.1.5 Robustness

To conclude this section, note that the choice of specification for the spatial model affects the results only marginally. The latter remain unchanged when the maximum distance threshold varies or when the concept of distance chosen is based on the closest neighbours (table 11.11 in Appendix 11.3.6). Lastly, the true value of parameter  $\rho$  does not affect the magnitude of the bias. Figure 11.5 shows that, at a given sampling rate, an estimate on a sample drawn by simple random sampling almost never makes it possible to find the true value of parameter  $\rho$ . In the case of a cluster survey, this value can be included in the confidence interval of the estimated parameter. However, the bias does not disappear when the magnitude or sign of this parameter vary. In any case, the bias mitigates the magnitude of the estimated spatial correlation.

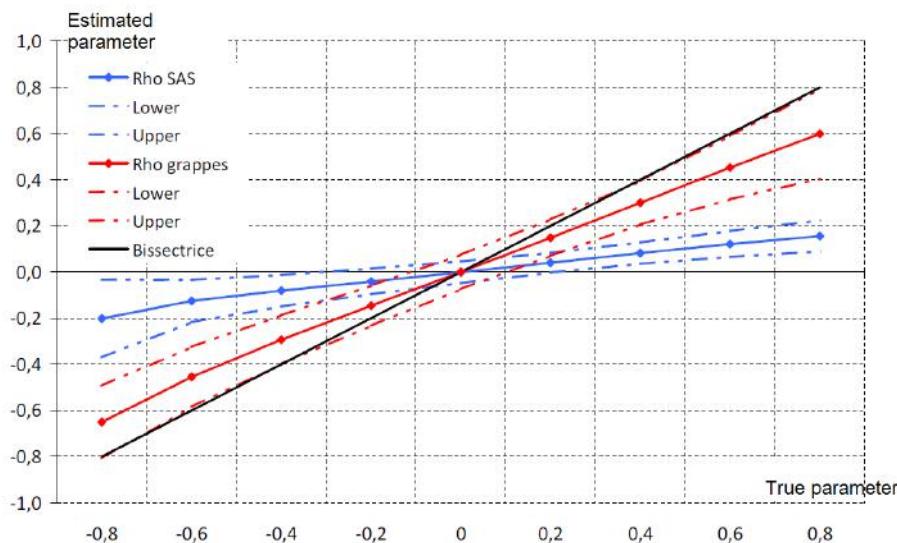


Figure 11.5 – Estimates of  $\hat{\rho}$  for various values of  $\rho$

**Note:** Full-line curves represent the estimated value  $\hat{\rho}$  based on the effective value  $\rho$  set for data simulation. The blue curve shows the case of data driven by simple random sampling and the red curve the case of cluster sampling. Dotted curves represent 95% confidence intervals for estimator  $\hat{\rho}$ .

Lastly, considering a SEM model (*Spatial Error Model*):  $Y_2 = X\beta + (1 - \lambda W)^{-1} \varepsilon$  does not radically affect the results (Table 11.12 in Appendix 11.3.6).

## 11.2 Prospects for resolution

One of the first positions one can adopt in the face of missing data is to ignore, consciously or not, this data and to directly apply the spatial model to the units observed. This mitigates the spatial correlation parameter relative to its true value, due to a "size effect" and a "sparsity effect".

The first effect comes from differences between the theoretical model and the estimated model regarding the dimension of the spatial weighting matrix. To remove it, the exhaustive data needs to be compared with the sample data according to a single geographical structure, and therefore on the same number of units. To compensate for the second, we need to be able to reconstruct the spatial correlations between observed and missing units. In this case, the location of units is always assumed to be known<sup>9</sup>.

In this section, we discuss the impact of two solutions commonly applied to empirical work. First, switching to a higher scale by aggregating data and second, imputing missing data. Both methods maintain the geographical structure of the data, but are more or less effective in reconstructing spatial correlations.

### 11.2.1 Moving to a higher scale by aggregation

In the absence of comprehensive individual data, much research has been carried out on an aggregated scale of regions, departments or employment areas. This choice depends crucially on the relevant scale of the economic issue, assumes the availability of a good estimator of the local average and can lead to an ecological bias (see Anselin 2002b for further details). Intra-zone correlations are then omitted, to the benefit of correlations between zones.

To evaluate this solution, we simulate 6,000 points drawn from a uniform distribution on a square space and assign them, as in section 11.1, values of  $X$  and  $Y$  according to a SAR data generating process characterized by a spatial autocorrelation parameter  $\rho = 0.5$ . These points are depicted on the left-hand surface of Figure 11.6. Then, this square space is divided according to a grid of size  $G \times G$  for different values of  $G$ , and to each centroid of each square is assigned the average of the points located inside this square. The centre and right panels of Figure 11.6 depict this configuration for  $G = 50$  and  $G = 20$  respectively.

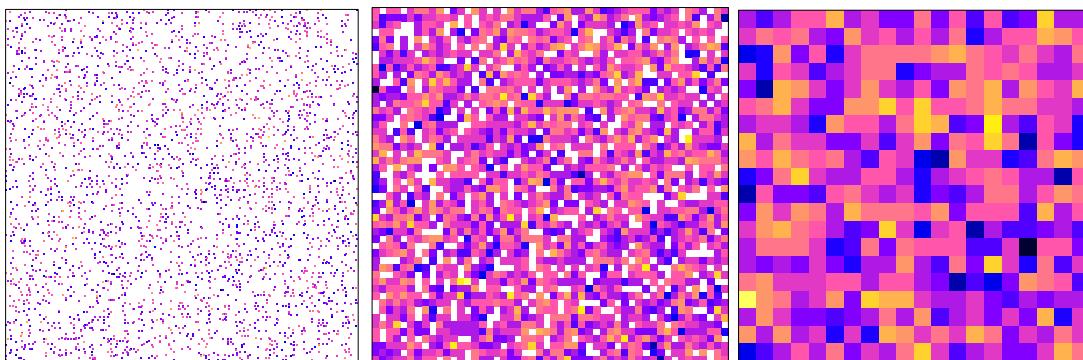


Figure 11.6 – Spatial data aggregation

**Note:** On the left panel, 6 000 items simulated from a uniform distribution, on the center panel, data aggregated in a grid of  $50 \times 50$  squares and on the right panel, data aggregated in a grid of  $20 \times 20$  squares

9. The lack of information on the location of certain units is another challenge for current research on spatial econometrics (Arbia et al. 2016) which exceeds the scope of this chapter.

The estimation of a SAR model on comprehensive data aggregated with  $G = 50$  provides a spatial correlation parameter  $\hat{\rho} = 0.47$  with standard deviation of  $\hat{\sigma}_\rho = 0.068$ . This parameter is significantly positive and the estimate includes 0.5 in its confidence interval. Aggregating data on squares would limit the loss of spatial interactions and minimize the bias in estimating the spatial correlation parameter. Figure 11.7 shows that parameters  $\rho$  and  $\beta$  are estimated precisely and without bias when the grid on which the data is aggregated is relatively fine. The finer the grid, the closer we are to the spatial structure of exhaustive individual data and therefore, the closer the spatial correlation is to its true value.

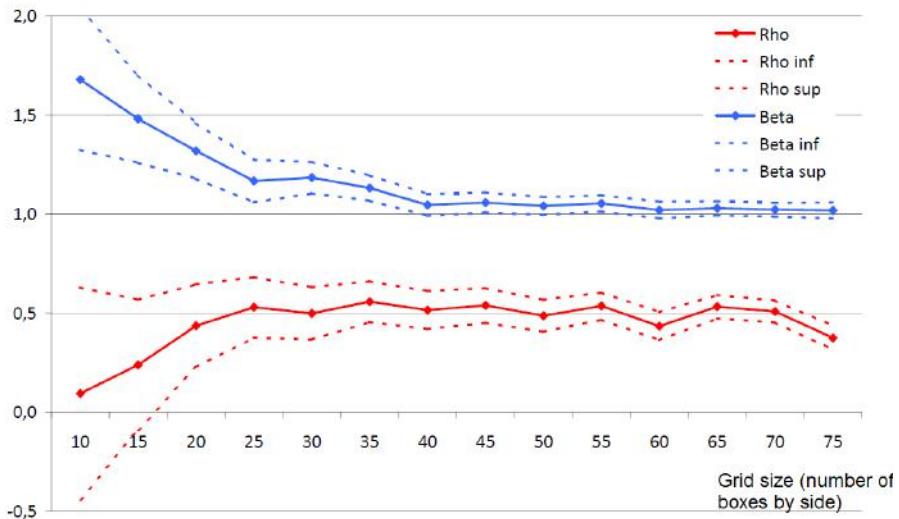


Figure 11.7 – Estimated parameters and fineness of the grid

**Note:** For each value of the size of the grid, the red curve displays the estimate  $\hat{\rho}$  and the blue curve the estimate  $\hat{\beta}$ . Dotted curves stand for the 95% confidence intervals. Results from simulating 6,000 points from a uniform distribution.

### Application to a sample

This procedure is replicated on data sampled by simple random sampling. The fineness of the grid fulfils the need for bias-variance arbitrage: fine squares reflect the distance between observations more accurately but lead to estimates of local averages that are less precise for each variable. Subject to assigning null weight and null values of the dependent and explanatory variables to squares without observation, we are able to identify the simulated spatial effect.

Table 11.4 shows the results of this procedure for different sample sizes and various spatial grids. In most cases, the true value of  $\rho$  is well within the confidence interval of the estimated parameter. For a small sample, an overly-coarse grid flattens out the spatial effects while an overly fine grid provides a poor estimate of individual variables. As before, the larger the sample, the more accurate the estimate.

These simulations tend to statistically validate the aggregation approach, provided that interpretation is not made directly at the individual level, but rests on strong hypotheses (coordinates of units drawn from a uniform distribution, homogeneous SAR process), rarely met in practice.

#### 11.2.2 Imputing missing data

To stay on the scale of the available data, the solution is to impute values to missing observations. This is another way of ignoring the "size effect": ensuring consistency between the spatial structure

n \ G	$\hat{\rho}$				$\hat{\beta}$			
	10	30	50	60	10	30	50	60
100	0.487*** (0.070)	0.494*** (0.060)	0.483*** (0.068)	0.478*** (0.072)	1.016*** (0.134)	1.007*** (0.115)	1.027*** (0.129)	1.030*** (0.134)
200	0.482*** (0.064)	0.499*** (0.045)	0.495*** (0.046)	0.489*** (0.050)	1.020*** (0.126)	0.998*** (0.084)	1.006*** (0.088)	1.011*** (0.095)
500	-0.093 (0.701)	0.488*** (0.032)	0.483*** (0.030)	0.489*** (0.032)	1.035*** (0.121)	1.022*** (0.060)	1.031*** (0.055)	1.021*** (0.059)
1000	-0.982 (0.159)	0.487*** (0.024)	0.485*** (0.020)	0.491*** (0.021)	1.048*** (0.119)	1.024*** (0.045)	1.028*** (0.038)	1.019*** (0.040)

Table 11.4 – SAR model estimated on data aggregated by squares of the grid

**Note:** Each line reflects the size  $n$  of the sample drawn from the 6,000 simulated points and each column reflects the fineness of the grid in terms of number of squares (a size 30 grid breaks the initial square in 900 boxes). \*\*\* denotes significance at 1%.

of survey data and administrative data. In the face of missing values in a survey or census, assigning "plausible" values to these units makes it possible to have a sample or even a complete population.

### Imputation methods

This section lists a number of traditional imputation methods. Interested readers may refer to a comprehensive handbook on survey theory, for example Ardilly 1994 or Tillé 2001, which provides more information, theoretical context, as well as other more advanced methods. In the case of imputation by ratio or by hot deck, explanatory variables  $X$  are assumed to be known exhaustively.

**Imputation by the mean.** The method of imputation by the mean (or by the median, or by the dominant class in the case of qualitative variables) is a common method consisting of replacing all missing values by the mean of the observed values. This method does not respect a possible econometric structure between different variables of the survey and may lead to false results in the estimation of such models.

**Imputation by ratio.** The attribution by ratio method involves mobilising the auxiliary information  $X$  available on the entire population, including the units for which the information of interest  $Y$  is missing, in order to impute plausible  $Y$  values. To do this, we assume the existence of a linear model  $Y = \beta X + \epsilon$ .  $\hat{\beta}$  is estimated by ordinary least squares, after which the value  $Y_{\text{ratio}} = \hat{\beta}X$  is imputed for the missing  $Y$ . The ratio of the  $Y$  over the  $X$ , in the case of quantitative data, is the same between the units observed and the units for which no information is available. This method can be refined by adding constraints to the units for which the estimate of  $\beta$  is calculated, for example on a specific domain or stratum.

**Hot deck imputation.** The hot deck method randomly connects a donor to a missing value, in contrast to the cold deck, which establishes this link deterministically. A donor here is an individual statistically "close" to the missing individual (they share similar values of auxiliary variable  $X$ , belong to the same stratum, to the same domain, or possibly are located in the same spatial position). The application of hot deck is based on the definition of a distance criterion, from which  $k$  neighbours of the valueless individual  $Y$  are determined. One individual is randomly selected from the  $k$  neighbours, uniformly or otherwise, to give its value to the new  $Y_{\text{hotdeck}}$ . Variants can be introduced, for example by limiting the number of times a single individual can be a donor, or by performing the hot deck sequentially.

We illustrate the proposed methods with a simple example. We simulate the geographic position

of  $N = 1,000$  points to which we assign variables  $X$  and  $Y$  following a SAR structure with  $\beta = 1$  and  $\rho = 0.5$ . We then draw samples by simple random sampling for different sizes  $n$ . For each sample, the  $N - n$  units not drawn are imputed by one of the methods mentioned above: imputation by the  $X$  ratio, imputation by statistical hot deck (neighbours have values close to  $X$ ) and imputation by geographic hot deck (neighbours are spatially close). Table 11.5 compares the results of these different methods to direct exploitation of sample.

$n$	Direct		Ratio		Statistical hot deck		Geographic hot deck	
	$\rho$	$\beta$	$\rho$	$\beta$	$\rho$	$\beta$	$\rho$	$\beta$
100	0.06 (0.06)	1.14*** (0.14)	0.05*** (0.01)	1.13*** (0.12)	0.03 (0.03)	1.06*** (0.13)	0.19*** (0.05)	0.12** (0.05)
200	0.09* (0.04)	1.10*** (0.08)	0.11*** (0.02)	1.11*** (0.08)	0.06** (0.02)	1.08*** (0.09)	0.22*** (0.04)	0.22*** (0.05)
500	0.19*** (0.02)	1.08*** (0.04)	0.25*** (0.02)	1.09*** (0.05)	0.19*** (0.02)	1.09*** (0.05)	0.31*** (0.03)	0.55*** (0.04)

Table 11.5 – Imputation methods

**Note:** Parameter  $\rho$  of SAR model, estimated by Monte Carlo, for a 100-size sample, after imputation by ratio, is 0.05, with empirical standard deviation of 0.01. \*\*\* indicates a 1% significance, \*\* 5% and \* 10%.

The choice of method has a significant impact on the results. Imputation by ratio seems to work well for both parameters even though it underestimates parameter  $\rho$  for small samples. On the other hand, the geographic hot deck method gives good results on the auto-correlation parameter but implies a very strong bias on parameter  $\beta$ . Finally, the statistical hot deck method seems to give similar results to the direct estimation on the sample. These results illustrate these methods on a very simple example and show their inability to find the initial parameters of the model.

#### Further reading

Imputation methods may bias estimates. In particular, the link between  $Y$  and  $X$  on which the imputation is based can be passed on to parameters estimated from a regression of  $Y$  on  $X$  (see Charreux et al. 2016 for a discussion on this point). Similarly, imputation methods can create a spatial structure *ex-nihilo* or, to the contrary, break the spatial correlations which are not taken into account.

Lastly, as mentioned in introduction, more refined methods of imputation using EM algorithms have been developed (LeSage et al. 2004; Wang et al. 2013a). However, they are complex, very specific to the type of information missing and still remain rarely applied.

### 11.3 Empirical application: the manufacturing sector in Bouches-du-Rhône

In order to illustrate the issues of estimating spatial models on survey data, we estimate a production function on plants from the manufacturing sector. The spatial approach makes it possible to measure the impact of interactions between the production processes of selected firms. Such *spillovers* between firms have already been highlighted by a significant literature on conurbation economies (see in particular LeSage et al. 2007, Ertur et al. 2007, López-Bazo et al. 2004, Egger et al. 2006).

### 11.3.1 Data

The SIRUS register (identification system in the register of statistics units<sup>10</sup>) lists all French firms, groups and their entities, contained in SIRENE (computerised system for the national register of enterprises and institutions<sup>11</sup>), the administrative register used for registering legal units. For each firm, information on turnover, main activity (accessible via the APE code (principal activity code), following the French nomenclature), total balance sheet, exports (administrative and full-time equivalent), physical address and list of plants is available.

Using the geographical information available (cadastral reference, road or city centre), the (x,y) coordinates of each entity have been successfully geolocated by the INSEE Methodology and Geographical Guidelines Division. This geographical data, combined with the economic data available in the SIRUS register, makes it possible to model econometric relations by taking into account the spatial interactions.

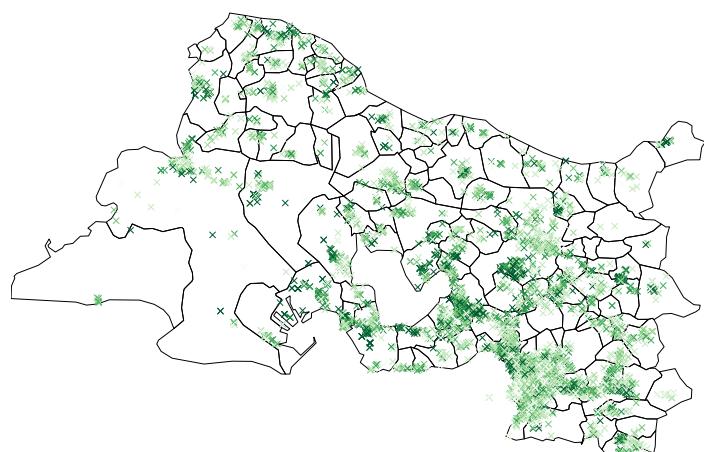


Figure 11.8 – Industrial companies in Bouches-du-Rhône

**Source:** SIRUS register, 2015

**Scope:** Companies of the manufacturing sector in Bouches-du-Rhône

**Note:** Darker green matches higher turnover

### 11.3.2 Identification

The production of a firm can be influenced by the geographical proximity of neighbouring companies. These interactions are designated by the concept of "externalities" which can be positive when the neighbourhood has a favourable impact on production (complementarities between sectors, integration of production chains, relationship with suppliers, transport, sharing of knowledge, etc.) or negative when they damage production (competition, pollution, traffic jams, etc.).

First of all, in order to carry out this analysis, it is necessary to choose coherent field, type of spatial links, and territory, so that the units considered maintain relations between them but not (or little) with the outside. In this case, we assume to have exhaustive data as regards the spatial effects on this territory. The choice to study the manufacturing sector in the Bouches-du-Rhône is didactic, but not meaningless. The presence of the port of Fos-sur-mer, the axis of the Rhône valley and the road nodes towards Toulouse and Italy make it a territory of interest (Figure 11.8 clearly illustrates the establishment of plants according to the transport networks). Of course, this estimate will not take national or even international trade relations into account, but will be exhaustive concerning local relations.

10. In French, *Système d'Identification au Répertoire des Unités Statistiques*.

11. In French, *Système Informatisé du Répertoire National des Entreprises et des établissements*.

The production level  $Y_i$  of a plant  $i$  can be stated according to a Cobb-Douglas production function:  $Y_i = AL_i^{\beta_L}K_i^{\beta_K}$ , where workforce  $L_i$  and capital  $K_i$  are input factors while  $A$  refers to general productivity of factors. Parameters  $\beta_L$  and  $\beta_K$  represent, respectively, the share of earned income and capital in production<sup>12</sup>. Traditionally, the term  $A$  refers to all mechanisms that influence production (human capital, technological progress, complementarities...) without being directly measurable. It can also be conceived of as representing the positive production externalities and be written:  $A = \exp(\beta_0) \prod_{j \in v_i} Y_j^{\rho \omega_{ij}}$ , where  $v_i$  refers to the neighbours of plant  $i$ , and  $Y_j$  the production level of a neighbouring plant  $j$ . The production function may be expressed in logarithm as:

$$\log(Y_i) = \beta_0 + \rho \sum_{j \in v_i} \omega_{ij} \log(Y_j) + \beta_L \log(L_i) + \beta_K \log(K_i) + \varepsilon_i \quad (11.3)$$

or:

$$\tilde{Y} = \beta_0 + \rho \mathcal{W} \tilde{Y} + \beta_L \tilde{L} + \beta_K \tilde{K} + \varepsilon$$

Terms with a tilda refers to the logarithm of these variables.  $\mathcal{W}$  is the spatial weighting matrix, such that  $\mathcal{W}_{i,j} = \omega_{ij}$ . This equation may be estimated using a SAR model. The parameter of spatial autocorrelation  $\rho$  captures the complementarities common to all units while  $\omega_{ij}$  captures specific complementarities resulting from the impact of the production of plant  $j$  on the production of plant  $i$ . The term  $\rho \omega_{ij}$  refers to the elasticity of the plant  $i$ 's production with respect to the production of plant  $j$ : when a plant  $j$  neighbouring  $i$  increases production by 1%, the production level of plant  $i$  increases by  $\rho \omega_{ij}\%$  through direct effects. Deriving equation 11.3 with respect to  $\log(Y_k)$ , we have:

$$\frac{d \log(Y_i)}{d \log(Y_j)} = \underbrace{\rho \omega_{ij}}_{\text{Direct effect from } j \text{ on } i} + \underbrace{\rho \sum_{k \neq j} \omega_{ik} \frac{d \log(Y_k)}{d \log(Y_j)}}_{\text{Indirect effect via } k}$$

Likewise, summing this expression over  $j$ ,  $\rho$  appears as the direct elasticity of plant  $i$ 's production with respect to the production of neighbouring plants:

$$\sum_j \frac{d \log(Y_i)}{d \log(Y_j)} = \rho + \rho \sum_j \sum_{k \neq j} \omega_{ik} \frac{d \log(Y_k)}{d \log(Y_j)}$$

### 11.3.3 Estimation

The equation 11.3 is estimated over 6,306 plants geolocated in Bouches-du-Rhône, belonging to the manufacturing sector<sup>13</sup>. This sector is particularly appropriate to a spatial estimate, since the geographical location is not directly part of its production activities (unlike trade, transport or agriculture), it is not too concentrated (as is the case in high technology) and does not make particular use of network logics other than spatial (contrary to finance or communications for instance).

Production  $Y_i$  from plant  $i$  is given by its turnover. The total balance sheet of the plant, which is a measure of its assets, is used as a proxy for capital  $K_i$ . These two variables, which are only available at firm level, are divided by the number of plants within the company. Lastly, workforce  $L_i$  is available at plant level in SIRUS.

Figure 11.8 shows the location of these plants. The intensity of the green in these cross-signs denotes their turnover: the darker the colour, the higher the turnover. Groups of plants with strong

12. These parameters can also be interpreted, respectively, as the elasticities of production with respect to work and capital.

13. The manufacturing sector encompasses companies whose main business belongs to Divisions 10 to 33 of NAF Rev 2. 2008.

turnover appear clearly, for example near Aix-en-Provence or around Fos-sur-Mer. As in the simulations in section 11.1, neighbourhood proximity is depicted by a weighting matrix based on distance. According to our definition, each plant has on average 109 neighbours, and 76 plants do not have neighbours<sup>14</sup>.

$\beta_0$	$\beta_L$	$\beta_K$	$\rho$
0.422	0.535	0.769	0.051
(0.050)	(0.015)	(0.009)	(0.009)

Table 11.6 – SAR model estimation: all plants

**Source:** *SIRUS register, 2015*

**Scope:** All plants from the manufacturing sector located in the Bouches-du-Rhône department, whose turnover and total balance sheet are strictly positive

**Note:** Estimated parameters are significant at 1%.

Table 11.6 shows the results of the SAR model estimated on all of plants within the manufacturing sector in Bouches-du-Rhône. Labour income and capital income shares are close to values generally estimated (roughly one-half to two-thirds for the latter, and one-third to two-thirds for the second, the high marginal return on capital being attributable to the choice of the industrial sector). There is a positive and significant spatial correlation: when the average turnover of the neighbouring units of plant  $i$  increases by 1%, the turnover of plant  $i$  increases by 0.05%.

### 11.3.4 Spatial estimates on samples

#### Sampling plans

As in Section 11.1, we replicate the model's estimate 11.3 on a sample of plants. Simple random sampling is used as a reference point, but is not common in the context of corporate surveys. Stratified sampling methods are more frequently employed in studies identifying the effect of the labour force and capital on turnover. These sampling methods have been presented in Section 11.1.2.

Stratification is carried out according to this workforce variable, assuming a correlation between headcount and turnover. Table 11.7 shows the strata created using a Neyman Allocation, based on dispersion of turnover within each of the strata. The dispersion within Stratum 4 is much higher than that of the other strata. As a consequence, we consider Stratum 4 to be exhaustive, *i.e.* we will always sample the 67 companies of Stratum 4 in order to limit the variance of estimation.

#### Results

In this section, we compare the results secured using a simple and stratified random sampling plan, varying the sample size:  $n \in \{250, 500, 1000, 2000\}$ .

Table 11.8 shows the parameters of the SAR model estimated from 1,000 draws by simple random sampling (*on the left*) and stratified sampling (*on the right*). In the case of simple random sampling, as well as in section 11.1, the traditional regression parameters  $\beta_L$  and  $\beta_K$  are correctly estimated. In contrast, spatial correlation parameter  $\rho$  is significant only for a sample size greater than 1,000 and always remains lower than the value it would take on full data.

The stratified survey plan applied to non-reweighted data skews traditional estimators  $\beta_L$  and  $\beta_K$  when regression is unweighted (Davezies et al. 2009). On the other hand, the bias on spatial

14. The units without neighbours, also referred to as "islands", do not participate in estimating the spatial correlation parameter  $\rho$ . The threshold is determined by a trade-off between minimising the number of neighbours and the number of islands.

Number of strata	Number of employees	Number of companies
1	0	3 628
2	1 to 9	2 742
3	10 to 99	770
4	100 and more	67

Table 11.7 – Constitution of strata

**Source:** *SIRUS register, 2013***Scope:** All plants from the manufacturing sector located in Bouches-du-Rhône, whose turnover and total balance sheet are strictly positive

n	Simple random sampling			Stratified sample		
	$\rho$	$\beta_L$	$\beta_K$	$\rho$	$\beta_L$	$\beta_K$
250	0.011 (0.021)	0.554*** (0.104)	0.768*** (0.078)	0.015* (0.009)	0.311*** (0.072)	0.813*** (0.056)
500	0.017 (0.016)	0.545*** (0.073)	0.773*** (0.052)	0.020** (0.008)	0.371*** (0.053)	0.796*** (0.041)
1000	0.024** (0.012)	0.542*** (0.051)	0.774*** (0.039)	0.024*** (0.007)	0.410*** (0.039)	0.793*** (0.029)
2000	0.034*** (0.010)	0.541*** (0.031)	0.770*** (0.023)	0.036*** (0.007)	0.457*** (0.028)	0.790*** (0.022)

Table 11.8 – Model 11.3 estimated on a random sample and on a stratified sample

**Source:** *SIRUS register, 2015***Scope:** All plants from the manufacturing sector located in Bouches-du-Rhône, whose turnover and total balance sheet are strictly positive

correlation parameter  $\rho$  appears less. This is because large companies likely to have significant spatial influence are all taken into account in the sample, due to this stratified survey plan.

The decision not to weight the regression is made by default. In traditional econometrics, it is of use to weigh observations before estimating an econometric model when the structure of the sampling plan is linked to the estimated variables. However, the question of using sampling weight as part of a SAR model has not been definitively addressed by the current literature<sup>15</sup>. In the current state of affairs, unweighted regression appears to be the safest and easiest choice. We do not explore this question further in this chapter.

### 11.3.5 Estimation on aggregated data

As discussed in section 11.2, an approach commonly used to circumvent the problem of missing data consists in moving to a wider scale by aggregating the sampled data. In order to move away from the administrative zonings, we divided the Bouches-du-Rhône department according to a  $G \times G$  grid (Figure 11.9).

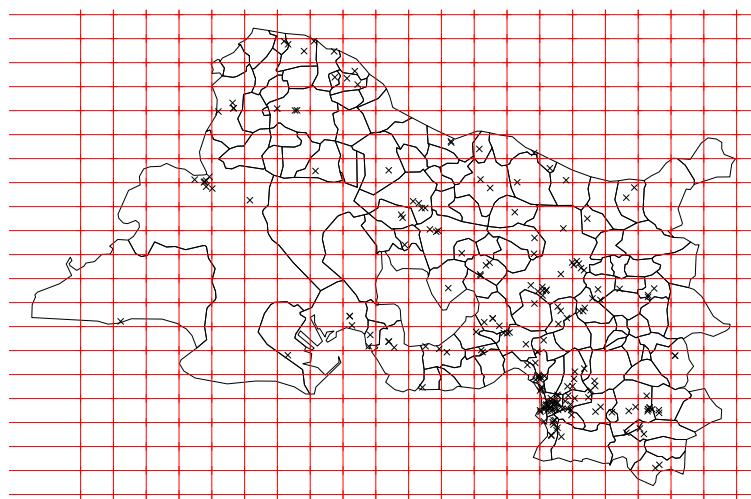


Figure 11.9 – Bouches-du-Rhône breakdown in a grid of  $20 \times 20$  squares

**Source:** SIRUS register, 2015

**Scope:** Plants from the manufacturing sector in Bouches-du-Rhône

From this grid, observations of a sample are averaged on each cell and the spatial analysis is carried out at the scale of the grid, with the distance being defined between the centroids of the cells. Null values are assigned to the variables and spatial weights of cells without observation, thus excluding them from the estimate without distorting the size of the spatial weighting matrix. Table 11.9 shows parameter  $\rho$  estimated for different sample sizes and various grid sizes.

Estimating spatial models based on aggregated data appears to circumvent the problem of missing data within a very simple context of data simulated uniformly across a territory. However, the application of this method to actual data is not straightforward. In particular, in this specific case, the parameter of spatial autocorrelation is still underestimated and is never significant. This could be due to the high concentration of plants in Bouches-du-Rhône, as the intra-cell distances are not, by definition, taken into account in this method. Spatial estimates on aggregated data thus require that the estimated phenomenon not be specific to a finer geographical scale.

15. For example, it is not clear whether it is necessary to involve the sampling weights in the spatial weighting matrix calculation  $\mathbf{W}$ ; this could also induce additional endogeneity, linked to the sample structure.

$n \backslash G$	20	30	50	60
100	0.007 (0.018)	0.009 (0.022)	0.014 (0.022)	0.015 (0.024)
200	0.013 (0.021)	0.007 (0.019)	0.015 (0.018)	0.018 (0.018)
500	0.024 (0.031)	0.023 (0.023)	0.012 (0.014)	0.013 (0.013)
1000	0.031 (0.026)	0.057* (0.040)	0.021* (0.015)	0.014 (0.012)

Table 11.9 – Parameter  $\rho$  estimated on aggregated data**Source:** SIRUS register, 2015**Scope:** Plants from the manufacturing sector in Bouches-du-Rhône**Note:** Parameters are estimated on a size- $n$  sample aggregated on a  $G \times G$  grid. For the sake of clarity, we show here only parameter values  $\rho$ .

### 11.3.6 Imputation of missing data

#### Implementation

The second approach, referred to in section 11.2.2, consists in imputing the missing data, *i.e.* to attribute estimated  $Y_i$  values to plants for which none are available. We consider three types of imputations for Bouches-du-Rhône plants: (*i*) *imputation by ratio*, which uses variables  $L$  and  $K$  representing workforce and capital as explanatory variables of the model, (*ii*) imputation by *statistical hot deck*, in the sense that the distance is calculated based on the values of  $L$  and  $K$ , that is to say that the neighbours of a plant are the plants that share similar staff and capital and (*iii*) imputation by *geographic hot deck*, where one is associated with an individual in the geographic sense.

The implementation of these techniques requires, in the first case, to estimate a linear model (function `lm` in R), and in the next two cases, to define the neighbours (function `knn` of package `class`), then randomly make a draw amongst them (function `sample` in R). These three approaches are tested on the manufacturing sector in Bouches-du-Rhône. 1,000 size- $n$  samples are drawn according to the principles of simple random sampling, after which the imputation process assigns values  $Y$  to the  $N - n$  companies not sampled. The results obtained are presented in table 11.10. As a reminder, the results on the entire population can be found in table 11.8.

#### Results

The results are highly variable, depending on the method used. Imputation by ratio makes it possible to maintain the linear structure between turnover, workforce and capital, resulting in unbiased and accurate estimates for coefficients  $\beta_L$  and  $\beta_K$ . On the other hand, estimation of  $\rho$  is very low, even more so than in the case of a direct estimation on a random sample (see table 11.8). Imputation does not take into account the spatial structure, which is deleted when the model is estimated on the data completed. Therefore, it is not relevant to apply spatial econometric models to data imputed using this method.

Imputation by statistical hot deck appears more promising. The estimators are on the right order of magnitude with respect to the resulting values for the population and are estimated accurately. A comparison with table 11.6 reveals a bias when  $\hat{\rho}$ ,  $\hat{\beta}_L$  and  $\hat{\beta}_K$  are estimated on small samples. Thus, the imputation by hot deck skews the estimators (Charreux et al. 2016) but enables the structure of the spatial correlations to be brought out. This is because the connection between donor

$n$	Ratio			Statistical hot deck			Geographic hot deck		
	$\rho$	$\beta_L$	$\beta_K$	$\rho$	$\beta_L$	$\beta_K$	$\rho$	$\beta_L$	$\beta_K$
250	0.002 (0.002)	0.560*** (0.112)	0.768*** (0.080)	0.043*** (0.009)	0.664*** (0.083)	0.646*** (0.059)	0.419*** (0.046)	0.028 (0.034)	0.104*** (0.023)
500	0.004 (0.003)	0.548*** (0.077)	0.774*** (0.058)	0.042*** (0.008)	0.613*** (0.061)	0.698*** (0.044)	0.412*** (0.035)	0.061* (0.034)	0.149*** (0.022)
1000	0.008** (0.003)	0.546*** (0.051)	0.774*** (0.037)	0.040*** (0.007)	0.577*** (0.040)	0.734*** (0.028)	0.389*** (0.035)	0.116*** (0.035)	0.217*** (0.023)
2000	0.017*** (0.004)	0.542*** (0.032)	0.773*** (0.024)	0.040*** (0.007)	0.562*** (0.031)	0.751*** (0.023)	0.333*** (0.022)	0.203*** (0.034)	0.338*** (0.022)

Table 11.10 – Imputation methods

**Source:** *SIRUS register, 2015*

**Scope:** Plants from the manufacturing sector in Bouches-du-Rhône, whose turnover and total balance sheet are strictly positive

and recipient implicitly maintains the structure of spatial interactions. It is also possible that the underlying spatial structure  $Y$  also exists for  $L$  and  $K$  and is recovered by imputation. Thus, the use of this imputation method for econometric analysis involves a trade-off between bias and variance on parameters  $\beta_L$  and  $\beta_K$ , traditional in sampling theory. However, in this instance, the method also offers the advantage of substantially reducing the pre-existing bias on  $\rho$ . These results, which are tested only on this data set and with a simple sampling plan, are to be used with caution. In any case, the effectiveness of this method is not based on the spatial proximity between the donor and the recipient, as the last example shows.

The imputation method using geographic hot deck results in aberrant estimations. Based directly on the spatial proximity between donor and receiver, it gives rise to very strong overestimation of the spatial effect (the estimated  $\hat{\rho}$  is greater than its value estimated on the full population), at the expense of the effect of other variables in the model (estimated  $\hat{\beta}$ 's are well below the values of these parameters estimated on the full population). In fact, according to this method, nearby spatial structures will have the same turnover  $Y$ , which creates *ex-nihilo* a very high positive spatial correlation. The use of the spatial dimension to remedy the problem of missing data is not for the immediate future. Table 11.13 in Appendix 11.3.6 shows the results achieved for a geographical hot deck imputation by limiting itself to plants with similar workforce. Parameter  $\rho$  is less overestimated but the results remain far removed from the estimate based on full data. It may be possible to use geographic information parsimoniously for imputation, but this would require more extensive analysis of the data set and good knowledge of its spatial structure.

## Conclusion

This chapter highlights the difficulties associated with the application of spatial econometrics to sampled data. There are two pitfalls in particular: (i) a "size effect" by which the estimate on a remote sample distorts the spatial weighting matrix, and (ii) a "sparsity effect" resulting from the omission of units spatially correlated with the units observed. Both effects tend to underestimate the magnitude of spatial correlations. However, the bias is lower in the case of a cluster survey and when the sample is larger.

Empirical studies typically resolve this problem by ignoring missing observations, aggregating data on a larger scale or imputing the missing values. The first solution is never desirable. The other two are far from perfect, because it is hard to rebuild a complex set of information from few

observations. The imputation by statistical hot deck is promising, but we do not demonstrate its validity in a general case.

While this issue is bound to become more important as social media and geolocated data become more prominent, estimating spatial models based on sampled data remains rare. For the time being, it is preferable to consider comprehensive data. This chapter warns against overly-expeditious solutions, such as aggregating data on a higher scale, calling upon simplistic imputation methods or removing any mention of missing data. When a relatively large sample is available, or derived from a cluster survey, a spatial estimate could then be considered, bearing in mind that the resulting spatial correlation parameter will probably be underestimated.

## Appendix

### Choice of model and neighbourhood matrix

Here, we consider the same procedure as in section 11.1 for different values of the parameters. Table 11.11 shows estimates comparables to those from Table 11.2 for different neighbourhood matrices. Table 11.12 displays results in the case of a Spatial Error Model (SEM) instead of a SAR model.

$n$	$\mathcal{M}$	$\rho$			$\beta$		
		2 neighbours	5 neighbours	Distance	2 neighbours	5 neighbours	Distance
50		0.020	-0.003	0.042	1.107***	1.050***	1.054***
		(0.110)	(0.172)	(0.043)	(0.115)	(0.095)	(0.125)
100		0.063	0.069	0.058*	1.112***	1.056***	1.054***
		(0.076)	(0.111)	(0.031)	(0.079)	(0.065)	(0.086)
150		0.097*	0.115	0.073**	1.107***	1.052***	1.049***
		(0.060)	(0.088)	(0.028)	(0.062)	(0.051)	(0.068)
250		0.150***	0.189**	0.101**	1.105***	1.050***	1.053***
		(0.047)	(0.065)	(0.026)	(0.049)	(0.040)	(0.052)

Table 11.11 – SAR Model - Monte Carlo Estimation

**Source:** *SIRUS register, 2015*

**Scope:** Plants from the manufacturing sector in Bouches-du-Rhône

**Note:** Standard deviations are in brackets.

$n$	$\mathcal{M}$	$\lambda$			$\beta$		
		2 neighbours	5 neighbours	Distance	2 neighbours	5 neighbours	Distance
50		-0.025	-0.110	0.008	1.003***	1.003***	1.002***
		(0.167)	(0.287)	(0.193)	(0.115)	(0.113)	(0.112)
100		0.008	-0.027	0.024	1.003***	1.004***	1.003***
		(0.113)	(0.182)	(0.124)	(0.080)	(0.078)	(0.078)
150		0.023	0.002	0.034	0.998***	0.998***	0.998***
		(0.090)	(0.144)	(0.099)	(0.065)	(0.063)	(0.063)
250		0.047	0.042	0.052	1.000***	1.000***	1.000***
		(0.069)	(0.108)	(0.079)	(0.051)	(0.050)	(0.050)

Table 11.12 – SEM Model - Monte Carlo Estimation

**Source:** *SIRUS register, 2015*

**Scope:** Plants from the manufacturing sector in Bouches-du-Rhône

**Note:** Standard deviations are in brackets.

### Imputation by stratified geographical hot deck

Table 11.13 provides the results using geographical hot deck imputation, restricted to plants with similar workforce, *i.e.* those of the same stratum (defined in table 11.7) as the plant with a missing value.

$n$	Geographic stratified hot deck		
	$\rho$	$\beta_L$	$\beta_K$
250	0.137 (0.037)	1.216 (0.100)	0.029 (0.026)
500	0.148 (0.031)	1.192 (0.077)	0.071 (0.025)
1000	0.156 (0.026)	1.121 (0.061)	0.149 (0.025)
2000	0.148 (0.019)	0.542 (0.048)	0.279 (0.025)

Table 11.13 – Imputation by geographic stratified hot deck

**Source:** *SIRUS register, 2015*

**Scope:** Plants from the manufacturing sector in Bouches-du-Rhône

## References - Chapter 11

- Anselin, Luc (2002b). « Under the hood : Issues in the specification and interpretation of spatial regression models ». *Agricultural Economics* 27.3, pp. 247–267.
- Arbia, Giuseppe, Giuseppe Espa, and Diego Giuliani (2016). « Dirty spatial econometrics ». *The Annals of Regional Science* 56.1, pp. 177–189.
- Ardilly, Pascal (1994). *Les techniques de sondage*.
- Belotti, F., G. Hughes, and A. Piano Mortari (2017a). « Spatial panel-data models using Stata ». *Stata Journal* 17.1, 139–180(42).
- Boehmke, Frederick J., Emily U. Schilling, and Jude C. Hays (2015). *Missing data in spatial regression*. Tech. rep. Society for Political Methodology Summer Conference.
- Burt, Ronald S. (1987). « A Note on Missing Network Data in the General Social Survey ». *Social Networks* 9, pp. 63–73.
- Charreux, C et al. (2016). « Econométrie et Données d’Enquête: les effets de l’imputation de la non-réponse partielle sur l’estimation des paramètres d’un modèle économétrique ».
- Chow, Gregory C. and An-Loh Lin (1976). « Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series ». *Journal of the American Statistical Association* 71.355, pp. 719–721.
- Cliff, A.D. and J.K. Ord (1972). *Spatial autocorrelation*. Pion, London.
- Cochran, William G (2007). *Sampling techniques*. John Wiley & Sons.
- Davezies, L. and X. D’Haultfoeuille (2009). *To Weight or not to Weight? The Eternal Question of Econometricians facing Survey Data*. Documents de Travail de la DESE - Working Papers of the DESE g2009-06. INSEE, DESE.
- Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). « Maximum likelihood from incomplete data via the EM algorithm ». *Journal of the royal statistical society* 39.1, pp. 1–38.
- Egger, Peter and Michael Pfaffermayr (2006). « Spatial convergence ». *Papers in Regional Science* 85.2, pp. 199–215.
- Ertur, Cem and Wilfried Koch (2007). « Growth, technological interdependence and spatial externalities: theory and evidence ». *Journal of Applied Econometrics* 22.6, pp. 1033–1062.
- Ferreiro, Osvaldo (1987). « Methodologies for the estimation of missing observations in time series ». *Statistics and Probability Letters* 5.1, pp. 65–69.
- Gile, Krista J. and Mark S. Handcock (2010). « Respondent-driven sampling: an assessment of current methodology ». *Sociological Methodology* 40.1, pp. 285–327.
- Goulard, M., T. Laurent, and C. Thomas Agnan (2013). « About predictions in spatial autoregressive models: Optimal and almost optimal strategies ». *Toulouse School of Economics Working Paper* 13, p. 452.
- Harvey, A. C. and R. G. Pierse (1984). « Estimating Missing Observations in Economic Time Series ». *Journal of the American Statistical Association* 79.385, pp. 125–131.
- Huisman, Mark (2014). *Imputation of missing network data*. Ed. by Reda Alhajj and Jon Rokne. Vol. 2. Springer, pp. 707–715. ISBN: 978-1-4614-6169-2.
- Jones, Richard H. (1980). « Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations ». *Technometrics* 22.3, pp. 389–395.
- Kelejian, H.H. and I.R. Prusha (2010). « Spatial models with spatially lagged dependent variables and incomplete data ». *Journal of geographical systems*.
- Koskinen, Johan H., Garry L. Robins, and Philippa E. Pattison (2010). « Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation ». *Statistical Methodology* 7.3", pp. 366–384.
- Kossinets, Gueorgi (2006). « Effects of missing data in social networks ». *Social Networks* 28.3, pp. 247–268.

- LeSage, James P., Manfred M. Fischer, and Thomas Scherngell (2007). « Knowledge spillovers across Europe: Evidence from a Poisson spatial interaction model with spatial effects ». *Papers in Regional Science* 86.3, pp. 393–421. ISSN: 1435-5957.
- LeSage, J.P. and R.K. Pace (2004). « Models for spatially dependent missing data ». *The journal of real estate finance and economics* 29.2, pp. 233–254.
- Little, Roderick J. A. (1988). « Missing-Data Adjustments in Large Surveys ». *Journal of Business and Economic Statistics* 6.3, pp. 287–296.
- Little, Roderick J. A. and Donald B. Rubin (2002). *Statistical analysis with missing data*. 2nd. Wiley, Hoboken.
- Liu, Xiaodong, Eleonora Patacchini, and Edoardo Rainone (2017). « Peer effects in bedtime decisions among adolescents: a social network model with sampled data ». *The Econometrics Journal*.
- López-Bazo, Enrique, Esther Vayá, and Manuel Artís (2004). « Regional Externalities And Growth: Evidence From European Regions ». *Journal of Regional Science* 44.1, pp. 43–73.
- Mardia, Kanti V. et al. (1998). « The Kriged Kalman filter ». *Test* 7.2, pp. 217–282.
- Pinkse, Joris and Margaret E. Slade (2010). « The Future of Spatial Econometrics ». *Journal of Regional Science* 50.1, pp. 103–117.
- Rubin, Donald B. (1976). « Inference and missing data ». *Biometrika* 63, pp. 581–592.
- Stork, Diana and William D. Richards (1992). « Nonrespondents in Communication Network Studies ». *Group & Organization Management* 17.2, pp. 193–209.
- Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies: cours et exercices avec solutions:[2e cycle, écoles d'ingénieurs]*. Dunod.
- Wang, W. and L.-F. Lee (2013a). « Estimation of spatial autoregressive models with randomly missing data in the dependent variable ». *The Econometrics Journal* 16.1, pp. 73–102.
- Zhou, Jing et al. (2017). « Estimating Spatial Autocorrelation With Sampled Network Data ». *Journal of Business and Economic Statistics* 35.1, pp. 130–138.



# 12. Small areas and spatial correlation

PASCAL ARDILLY

INSEE

PAUL BOUCHE

ENSAI - Sciences Po

WENCAN ZHU

ENSAI

---

<b>12.1 Setting up the model</b>	<b>306</b>
12.1.1 Background and objectives . . . . .	306
12.1.2 Standard individual linear model . . . . .	307
12.1.3 Individual linear model with spatial correlation . . . . .	308
12.1.4 How to deal with qualitative variables using a generalised linear mixed individual model . . . . .	310
12.1.5 Extension to models defined at area level . . . . .	311
<b>12.2 Forming the "small area" estimator</b>	<b>312</b>
12.2.1 BLUP estimation strategy: the standard individual model . . . . .	313
12.2.2 Application to the individual linear model with spatial correlation . . . . .	315
12.2.3 Application to the Fay and Herriot model . . . . .	315
12.2.4 Strategy for non-linear models . . . . .	316
<b>12.3 The quality of estimators</b>	<b>316</b>
12.3.1 An iterative process . . . . .	317
12.3.2 The problem of bias . . . . .	317
12.3.3 Mean square error . . . . .	320
<b>12.4 Implementation with R</b>	<b>320</b>

---

## Abstract

When we want to circulate the results of a survey on small populations, particularly if we are dealing with small geographical areas, the low sample size matching these populations can lead to estimates that are not accurate enough. The classical sampling theory does not provide a satisfactory solution to this problem and specific estimation techniques must therefore be used, based on using auxiliary information and on models of varying complexity. All these models are a formal link between the variable studied and auxiliary variables. The simplest form is a linear link but there are other non-linear models (Poisson model, logistics model). Most of the models isolate local area-specific effects. Correlations between these effects can be introduced, all the stronger given that the areas are geographically close. This spatial correlation is then likely to improve the quality of localised estimates.

This chapter is devoted to a general introduction to the issue known as the "small-area estimation", with particular attention to considering spatial correlation in the models.

## 12.1 Setting up the model

### 12.1.1 Background and objectives

Survey statisticians have a particular interest in estimating unknown  $\theta$  parameters, defined in a finite and generally large population. Most parameters are totals or immediate derivatives of totals such as means or proportions. More rarely, we find non-linear functions that can still be expressed as total functions (ratios, variances in the population, correlation or regression coefficients). Depending on the survey topic, we may also want to estimate highly non-linear parameters, such as quantiles or inequality indicators, which are not written as total functions.

The parameters are defined using one (or more) variable(s) of interest and are formally stated using expressions generally involving all individuals in population  $U$ . We will designate  $Y$  as the variable of interest, which will subsequently be considered as unique. The individuals of  $U$  are identified by index  $i$ , and if parameter  $\theta$  is a total  $T$  then  $T = \sum_{i \in U} Y_i$ .

When individual value  $Y_i$  for each individual  $i$  of  $U$  is not available, we can estimate  $T$  by sampling, *i.e.* from information  $Y_i$  obtained from a responding sample, designated  $s$ , included in  $U$ . The sample is usually drawn from a complex sampling design, for example combining stratification, unequal probability sampling and several sub-samples. Some parameters of interest are not defined for the whole population  $U$  but on a sub-population, designated  $d$ . Such a sub-population is called a *domain* (or an *area*), and we then have to deal with an area estimate. In this case, the parameter of interest  $\theta$  may be the total over the area, *i.e.*  $T_d = \sum_{i \in d} Y_i$ , which must be estimated from collected data. The classical sampling theory assigns each sampled unit  $i$  a sampling weight  $w_i$ , a positive real coefficient that depends on the sampling method and how non-responses are handled, which "expands" the value of the variable of interest  $Y_i$ . To estimate a defined total  $T$  over the complete population  $U$ , the estimator takes a linear form  $\hat{T} = \sum_{i \in s} w_i Y_i$ . To estimate a total over an area  $d$ , we simply restrict the sum to the elements of  $d$  without modifying their weighting, *i.e.*  $\hat{T}_d = \sum_{i \in s \cap d} w_i Y_i$ . If parameter  $\theta$  is a mean over  $d$ , now designated  $\bar{Y}_d$  (including proportions, which are means of Boolean variables), we estimate the size  $N_d$  of the area using  $\hat{N}_d = \sum_{i \in s \cap d} w_i$  (a size is a total of constant individual values equal to 1) and we form ratio  $\hat{\bar{Y}}_d = \hat{T}_d / \hat{N}_d$ . But if we know  $N_d$ , we can also use the alternative estimate  $\hat{\bar{Y}}_d = \hat{T}_d / N_d$ .

In all cases, sampling leads to a specific error of estimators  $\hat{T}_d$  and  $\hat{\bar{Y}}_d$ , summarised by means of two indicators respectively called *bias* and *sampling variance*. Consider the case of  $\hat{\bar{Y}}_d$ . The bias means the difference between the expected value of  $\hat{\bar{Y}}_d$ , *i.e.* the expected "on average" estimate given the uncertainty that leads to the creation of  $s$ , and the  $\bar{Y}_d$  parameter, while sampling variance measures the sensitivity of the estimate  $\hat{\bar{Y}}_d$  to responding sample  $s$ . A precise sampling design results in a low bias and a low variance. The estimators derived from the sampling theory and used by survey statisticians are generally bias-free or have negligible bias. The sampling variance is a decreasing function of  $n_d$ , where  $n_d$  is the size of the responding sample matching area  $d$ , *i.e.* size of  $s \cap d$ . When  $n_d$  is small enough that the quality targets for estimate  $\hat{\bar{Y}}_d$  are not reached, there is a *small area estimation problem*.

To address this difficulty, when it is no longer possible to increase the value of size  $s$ , designated  $n$ , we have to create a new theoretical context to make the final estimate of parameter  $\theta$  (total  $T_d$  or mean  $\bar{Y}_d$ ) less sensitive to responding sample  $s$  (or  $s \cap d$ , which is equivalent). This is done using a *modelling* technique. It means putting oneself within a hypothetical framework that simplifies reality (this is the general definition of a model). The usual approach is to consider that  $Y_i$  is

essentially explained by a set of known individual variables  $X_i$  for each unit  $i$  of the population while involving a few  $\delta$  *a priori* unknown quantities - parameters of the model. It will be enough to estimate these quantities  $\delta$  to be able to deduce any unknown value  $Y_i$  (corresponding to cases  $i \notin s$ ), and therefore *ultimately* the value of parameter  $\theta$ .

The use of modelling essentially requires auxiliary information to be available. Of course, we are thinking of variables known at individual level over the entire population  $U$ . Let's assume that the auxiliary information about individual  $i$  consist in  $p$  individual variables, designated  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$ , and start from the principle that there is a "sufficiently reliable" link between such values and the variable of interest  $Y_i$ . This link is by construction considered valid when applied to the entire population  $U$ , without knowing anything other than  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$ . It must remain valid if we are limited to the responding sample  $s$ , which means that the information provided by belonging to the responding sample should not lead the statistician to change the formal expression of this relationship (so-called "uninformative" sampling design). In an ideal world where everything is simple, there would be a certain function  $f$  such that for any individual  $i$  of  $U$  we have  $Y_i = f(X_{i,1}, X_{i,2}, \dots, X_{i,p}; \delta)$  where  $\delta$  is a vector parameter unknown at this stage, known as a parameter of the model. In this perfect context, the functional form of the  $f$  function is fully known but it is nonetheless configured by  $\delta$ . If, using information collected during the survey, we manage to estimate the  $\delta$  parameter satisfactorily, we will be able to predict the values  $Y_i$  of all individuals  $i$  not sampled (or sampled but not responding) and therefore predict  $\theta$ .

The traditional framework of sampling statistics is that the sampling theory is not based on any modelling and considers that the variable of interest  $Y$  is not random (it is therefore deterministic). It is the sample selection procedure and the non-response mechanism that introduce uncertainty and this uncertainty allows any estimator, such as mean estimator  $\bar{Y}_d$ , to be considered as a random variable. However, if some mathematical modelling describe  $Y$ , since the reality is not that of an ideal and simple world, it would not be reasonable to assume that there is equality between  $Y_i$  value and any value of the type  $f(X_{i,1}, X_{i,2}, \dots, X_{i,p}; \delta)$ , because the relationship between  $Y_i$  and  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$  would be too restricted and therefore not credible. Therefore, function  $f$  must be considered as including a random component  $U_i$ , the first characteristic of which is to be guided by chance. We should now abandon the traditional environment of the sampling theory and consider that the  $Y$  variables are random variables, such as  $Y_i = f(X_{i,1}, X_{i,2}, \dots, X_{i,p}, U_i; \delta)$ .

### 12.1.2 Standard individual linear model

The structure of the model therefore uses specific and explicit uncertainties that have no relation to sampling uncertainty. In certain circumstances, it is customary to introduce an individual random variable  $U_i$  which is zero on average and thus linked to  $Y_i$ , for all  $i$  of  $U$  (Equation 12.1) :

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_p X_{i,p} + U_i \quad (12.1)$$

Auxiliary variables  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$ , which are perfectly deterministic, are called "fixed effects". The uncertainty of the model relating to  $Y_i$  values should not be confused with the sampling uncertainty that determines the sample composition  $s$ . At this stage, the special features of the context of small area estimation are revealed. As population  $U$  is partitioned into  $D$  areas, we consider that if  $i$  belongs to area  $d$ , uncertainty  $U_i$  - zero *on average* - is composed of an effect (random) specific to area  $d$ , designated  $\tau_d$ , and an individual (random) residual designated  $e_i$ . We therefore have:

$$U_i = \tau_d + e_i. \quad (12.2)$$

In the simplest approach, the two components  $\tau_d$  and  $e_i$  are assumed to be independent, the  $\tau_d$  are mutually independent, just as the  $e_i$  are mutually independent. The expected value and variance associated with the model's uncertainty will be designated  $\varepsilon$  and  $v$ , so the simplest assumptions supporting this model are:

- for expected values,  $\varepsilon(\tau_d) = 0$  and  $\varepsilon(e_i) = 0$ ;
- for variances,  $v(\tau_d) = \sigma_\tau^2$  and  $v(e_i) = \sigma_e^2$ .

Furthermore, all possible covariances involving these elementary components are zero. So, overall  $\varepsilon(U_i) = 0$  and  $v(U_i) = \sigma_\tau^2 + \sigma_e^2$ . The format of this model makes it possible to create a correlation between the variables of interest associated with units of the same area since  $\forall i \in U$ ,  $\forall j \in U$ ,  $j \neq i$ . If  $i \in d$  and  $j \notin d$  then  $cov(Y_i, Y_j) = cov(U_i, U_j) = 0$  and if  $i \in d$  and  $j \in d$  then  $cov(Y_i, Y_j) = cov(U_i, U_j) = \sigma_\tau^2$ . Thus, the variances-covariances matrix of the vector of the  $Y_i$ , where  $i$  covers  $U$ , has the form of one diagonal matrix per block, as each block is associated with an area and can be described using a diagonal including everywhere  $\sigma_\tau^2 + \sigma_e^2$  while all the other elements of the block take the constant value  $\sigma_\tau^2$ .

Due to assumptions covering the moments of random components, such a model can only strictly be applied to quantitative and continuous variables  $Y$  - which in particular excludes any qualitative variable of interest (and therefore parameters defined as proportions). Random effect  $\tau_d$  is a local effect interpreted as being the component of the variable of interest explained by belonging to the area beyond the information contained in individual variables  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$ . The areas are often geographical areas and  $\tau_d$  intends to reflect the purely explanatory part due to the geographical location of the unit. Assessing the true explanatory part of the location over a given area, and even defining a geographical effect, is a rather philosophical question. Indeed, because it is an easy and practical explanation, one can always consider a significant residual effect that would be due to inadequate consideration of the truly explanatory individual auxiliary variables as a geographical effect. In other words, if there are geographical elements that explain  $Y$ , they should ideally be translated in one way or another into fixed effects vector  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$ . Therefore, *a priori*, we have to see local effect  $\tau_d$  as an "interference" effect and seek to minimise its importance. The smaller parameter  $\sigma_\tau^2$ , *i.e.* the weaker  $\tau_d$  numerical values, the more the explanatory nature will be based on fixed effects  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$  and therefore the better the model. As the covariance structure is complex, we say that the model belongs to the family of general linear models.

With such a model, the expected value of random variable  $Y_i$  is a linear function of the  $\beta$  parameters. The largest explanatory component of  $Y_i$  consist in non-random effects  $X_{i,j}$  (fixed effects) however the residual component  $\tau_d$  attributed exclusively to the area is random (random effect). For these reasons, we are talking about *linear mixed model*.

If we use the designations in section 12.1.1, we confirm that  $Y_i = f(X_{i,1}, X_{i,2}, \dots, X_{i,p}, U_i; \delta)$ , where vectorial parameter  $\delta$  brings together all the unknown quantities appearing in the model, namely  $\delta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p, \sigma_\tau^2, \sigma_e^2)$ . It has a dimension  $p+3$ , distinguishing  $p+1$  actual parameters associated with explanatory fixed effects and two real parameters associated with the variances-covariances structure attached to the model.

### 12.1.3 Individual linear model with spatial correlation

The standard linear mixed model expresses the assumption of zero correlation between uncertainties  $U_i$  associated with individuals belonging to two separate areas. This situation is not necessarily credible, because there is no reason why the limits of geographical zoning making up the areas should be a barrier that suddenly stops any propagation of the measured phenomena. In general, there is a form of natural spatial continuity of the behaviours of localised individuals and two geographically-close individuals on the ground are more likely to display similar  $Y$  values

than two distant individuals. From this viewpoint, a relationship between the geographical effects characterising nearby areas seems quite natural.

From a technical viewpoint, we can try to reflect this situation by introducing a correlation that only considers the distance between areas. The analytical form of the correlation is free, provided it decreases when distance increases. In this spirit, we can rely on a model that exactly keeps the expressions 12.1 and 12.2 but that ensures  $\forall i \in d, \forall j \in d'$ , if  $i \neq j$ :

$$\text{cov}(Y_i, Y_j) = \text{cov}(U_i, U_j) = \text{cov}(\tau_d, \tau_{d'}) = \sigma_\tau^2 \exp\left(-\frac{1}{\rho} \text{dist}(d, d')\right) \quad (12.3)$$

where  $\text{dist}(d, d')$  is a defined distance between areas  $d$  and  $d'$ . For example, we can take the normal Euclidean distance calculated from the coordinates of the centroids of the two areas involved. Coefficient  $\rho$  is a scale parameter that allows better adjustment of the model. The more the distance influences covariance, the closer  $\rho$  will be to zero. In the particular case where  $d = d'$ , and when  $i \neq j$ , then  $\text{cov}(Y_i, Y_j) = \text{cov}(\tau_d, \tau_d) = \sigma_\tau^2 \exp(0) = \sigma_\tau^2$ . If  $i = j$ , the variance of the individual effect is added, i.e.  $\text{cov}(Y_i, Y_j) = \text{cov}(U_i, U_j) = \sigma_\tau^2 + \sigma_e^2$ . This time, the variances-covariances matrix is a full matrix, without zeros. Nonetheless we can consider, as an interesting variant, that the distance becomes infinite when it exceeds a certain threshold. This allows many zeros to be reintroduced into the matrix, so facilitating subsequent digital processing (especially by saving random access memory). Under such circumstances, there are slightly more model parameters since new parameter  $\rho$  must be taken into account, so  $\delta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p, \rho, \sigma_\tau^2, \sigma_e^2)$ .

Another approach is to introduce a simple relationship between local effects  $\tau_d$  of the different areas, ensuring that this relationship is all the stronger as the areas are closer together. Thus, we can consider that the local effect associated with a given area is "almost" a linear combination of local effects of the areas surrounding it, with a linking intensity that diminishes as we move away from the given area. The intensity of the link between the effects  $\tau_d$  is reflected by two elements, on the one hand a system of coefficients  $\alpha_{d,d'}$ <sup>1</sup> that govern the relative influence of the different areas distinguished  $d'$  over a given area  $d$ , on the other hand a parameter  $\rho$  between -1 and 1 which governs the absolute value of the linking intensity. For all  $i$ , we state:  $\sum_{d'=1, d' \neq d}^D \alpha_{d,d'} = 1$ . The proposed relationship between random effects is  $\tau_d \approx \rho \sum_{d'=1, d' \neq d}^D \alpha_{d,d'} \tau_{d'}$ . In matrix writing, this becomes:

$$\begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_D \end{pmatrix} = \rho \cdot \begin{pmatrix} 0 & \alpha_{1,2} & \dots & \alpha_{1,D} \\ \alpha_{2,1} & 0 & \dots & \alpha_{2,D} \\ \vdots & \ddots & 0 & \vdots \\ \alpha_{D,1} & \dots & \dots & 0 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_D \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_D \end{pmatrix} \quad (12.4)$$

by introducing an uncertainty vector  $u_d$  that follows a Gaussian distribution with variance  $\sigma_u^2 I_D$ . This is known as the SAR model (*Simultaneous AutoRegressive model*).

Arbitration between the latter method and the former one is not obvious *a priori*, which is why the only advice to be offered at this stage is to test both methods and then use the available quality assessment tools, particularly those mentioned in part 12.3.

1. Parameters  $\alpha_{d,d'}$  are the weightings  $w_{d,d'}$  of weighting matrix  $W$  used in the previous chapters. In this chapter,  $w$  means sampling weighting.

The introduction of a spatial correlation in the basic linear mixed model does not change any of the restrictive use conditions. Such a model can only be used to estimate  $\theta$  parameters built from a quantitative and continuous variable of interest. Moreover, it loses much of its benefit if the areas are geographically large because the distance considered is measured between the centroids of the areas.

In practice, to limit the number of non-zero coefficients in the variance-covariance matrix of local effects (and thus speed up the calculations and/or insufficient memory problems), we completely neutralise influence  $\alpha_{d,d}$  of  $d$  areas located beyond a certain distance of  $d$ , or even possibly not in the immediate vicinity of reference area  $d$ . Nonetheless, it is difficult to avoid the problems posed by "edge effects" that arise when an area is on the edge of a larger territory, because all its neighbours cannot be taken into account. For example, this is almost systematically true for the border territories of states.

#### 12.1.4 How to deal with qualitative variables using a generalised linear mixed individual model

##### The logistic model

The parameters for counting any sub-population are based on individual qualitative variables. Let us assume that we want to estimate the total number of individuals  $\theta$  confirming a given property  $\Gamma$  - such as "being a woman" or "being a farmer under 50". If we define individual variable  $Y_i = 1$  when  $i$  confirms  $\Gamma$  and otherwise  $Y_i = 0$ , it is easy to confirm that  $\theta = \sum_{i \in U} Y_i$ . Random variable  $Y$  defined in this way is a dummy variable that quantifies an initially-qualitative individual piece of information. By dividing  $\theta$  by size  $U$ , we get the proportion of individuals in the population who confirm property  $\Gamma$ . Unfortunately, model 12.1 is not at all appropriate for this type of variable. We work around the difficulty by opting for a modelling fully compatible with the dummy variables where Bernoulli distribution will provide the distribution of  $Y_i$ . This is a distribution that loads the value 1 with a probability  $P_i$  and value 0 with a probability  $1 - P_i$ . We can therefore consider that for any individual  $i$  from the overall population  $U$ , variable  $Y_i$  is a random variable that obeys Bernoulli distribution  $\mathcal{B}(1, P_i)$ . The core of the model follows as we will link the  $P_i$  parameter to the individual characteristics of  $i$  summarised by auxiliary variables  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$  and we will introduce a local random effect  $\tau_d$ . The functional form that links  $P_i$  to  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$  and to  $\tau_d$  must be compatible with the constraint  $P_i \in [0, 1]$ . There are various options, but the most common one is to state, for all  $i$  in  $d$ :

$$\log \left( \frac{P_i}{1 - P_i} \right) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_p X_{i,p} + \tau_d \quad (12.5)$$

We are talking about a *logistic model*. The expected value of random variable  $Y_i$  is  $P_i$ , which is obviously not a linear function of the  $\beta$  parameters (unlike in 12.1). For this reason, we say that the model represented by Equation 12.5 is a *generalised linear mixed model*. The class of models taking the form in 12.5 distinguishes between models where local effects  $\tau_d$  are mutually independent, as in 12.2, and models with spatial correlation, as in 12.3 or 12.4.

##### The Poisson model

Qualitative information is sometimes aggregated when handling statistical units. If we take the previous example, in the case where the units are households and no longer physical individuals, for each household  $i$  we have the total number of individuals  $Y_i$  confirming property  $\Gamma$  (the number of women in the household, or the number of farmers under the age of 50 in the household). This variable is no longer a dummy variable but a variable that can take any integer value (in practice this value has always an upper bound). Under these conditions, the Poisson distribution is a fairly

simple natural distribution that can be associated with  $Y_i$ . It has a single real parameter  $\lambda_i$  (strictly positive) that will be made to depend on unit  $i$  through individual characteristics  $X_{i,1}, X_{i,2}, \dots, X_{i,p}$  and a local random effect  $\tau_d$ . The  $\lambda_i$  parameter is often transformed using a simple function before being linked to explanatory factors. In practice, the logarithm function is mainly used, meaning that the complete - generalised linear mixed - model is expressed as follows:

$$\begin{aligned} Y_i &\sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_p X_{i,p} + \tau_d. \end{aligned} \quad (12.6)$$

Once again, local random effects  $\tau_d$  can be considered as mutually independent, as in 12.2, or spatially correlated, as in 12.3 or 12.4.

### 12.1.5 Extension to models defined at area level

#### The Fay and Herriot model

Taking sampling into account, we can produce estimators of any parameter, particularly totals  $T_d$  (or means  $\bar{Y}_d$ ) defined at area level  $d$ . These estimators are constructed with the individual sampling weights  $w_i$  (themselves a function of the sampling method used). They only use information about area  $d$ , which is why they are called *direct estimators*. It is possible to construct a model based on these estimators, designated  $\hat{T}_d$  for totals and  $\hat{Y}_d$  for means. The statistical unit modelled is then no longer the individual but the area. The aim is to link the available information  $\hat{T}_d$  or  $\hat{Y}_d$  to a set of explanatory variables, these being adapted to the level handled. Of course, they must characterise the areas and no longer the individuals. Local effects  $\tau_d$  retain their nature and interpretation, just as in Equation 12.2.

A famous model is the so-called *Fay and Herriot* model, part of the family of linear mixed models. If the explanatory variables selected at domain level are designated  $X_{d,1}, X_{d,2}, \dots, X_{d,p}$ , the most basic version of the model is written:

$$\bar{Y}_d = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + \dots + \beta_p X_{d,p} + \tau_d. \quad (12.7)$$

The variable explained here is the true mean in area  $d$ . Since this value is unknown, a step must be added to replace it with an estimate. At this stage, estimate  $\hat{Y}_d$  made from the survey is certainly not good quality since sample  $s \cap d$  is small, nonetheless it exists and can be linked to the true value by introducing an error term  $err_d$  according to

$$\hat{Y}_d = \bar{Y}_d + err_d. \quad (12.8)$$

Variable  $err_d$  is the sampling error. This last equation has nothing to do with a model, it is simply the definition of sampling error. Generally, estimator  $\hat{Y}_d$  is weighted so as to be unbiased or have negligible bias (if there was a calibration, for example, and however if we consider that the non-response was correctly handled) so that the sampling error has expected value zero when taking sampling uncertainty into account, *i.e.*  $E(err_d) = 0$ . The variance of the error depends on sampling but we know it varies as the inverse of  $n_d$ . From now on, we will designate this variance  $\psi_d$ . Combining the two previous equations leads to the operational formula:

$$\hat{Y}_d = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + \dots + \beta_p X_{d,p} + \tau_d + err_d. \quad (12.9)$$

As we saw with the individual models, we can make an assumption of independence between local effects  $\tau_d$  or instead propose spatial correlation, structured as in equations 12.3 or 12.4.

As assumptions on the expected value and variance of effects  $\tau_d$  are strictly speaking only compatible with true means  $\bar{Y}_d$  that have continuous distributions, it is needless to say that the individual variable of interest  $Y_i$  collected for individuals should be quantitative and continuous. Given that, if the variable of interest  $Y_i$  is qualitative and if area  $d$  has a large enough size  $N_d$ , we can consider - sometimes a little boldly! - than true mean  $\bar{Y}_d$  may *a priori* take a large enough number of values for this set to be considered as continuous, *i.e.* without "hole". Size  $N_d$  is the essential parameter. For example, consider  $Y_i$  the indicative variable characterising the "woman" condition. Mean  $\bar{Y}_d$  is then the proportion of women in the population of the area. If  $N_d = 10$ , this mean can take values  $k/10$ , where  $k$  is an integer between 0 and 10, which is a long way from creating a "continuous" situation. If  $N_d = 10000$ , the mean can take values  $k/10\,000$ , where  $k$  is an integer between 0 and 10 000, which makes the assumption of continuity much more plausible. This is why we can conclude that model 12.9 is acceptable for estimating proportions (qualitative variables of interest) when areas  $d$  are not too small.

### The Poisson model

Although the Fay and Herriot model adapts well to qualitative variables, *i.e.* parameters that are defined as proportions by area of individuals confirming a property  $\Gamma$  (similar to a sub-population  $\Gamma$ ) or as headcount by area of these same individuals, under certain circumstances it may be preferable to have a model more specifically adapted to the counts. Designate  $N_{\Gamma,d}$  as the total number of individuals in the area  $d$  belonging to sub-population  $\Gamma$ . The sample is used to form the unbiased (or nearly) estimator  $\hat{N}_{\Gamma,d} = \sum_{i \in s \cap d \cap \Gamma} w_i$ . This estimator only uses area-related information, so it is a direct estimator, and it is poor quality since sample  $s \cap d$  is small. Nonetheless, it is a calculable random variable for which the distribution can be modelled using a Poisson distribution. This distribution, which is dependent on a single real parameter  $\lambda_d$ , a function of the area, is particularly suitable for counts. We can show that  $\lambda_d$  is the expected value of  $\hat{N}_{\Gamma,d}$  and it should therefore be numerically quite close to this estimate. At this stage, this is a first assumption and not a property that would arise from the sampling theory. Nonetheless, the risk taken remains small because the asymptotic behaviour of direct estimators is close to a Gauss distribution, which the Poisson distribution itself is close to if its parameter is large enough.

The core of the model comes from the following. We generally consider that the logarithm of the  $\lambda_d$  parameter is written in this way

$$\log(\lambda_d) = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + \dots + \beta_p X_{d,p} + \tau_d \quad (12.10)$$

using the designations in the previous sections. Random variable  $\tau_d$  keeps the same interpretation. The aim is to distinguish the effect of the location of statistical units beyond what fixed effects  $X_{d,1}, X_{d,2}, \dots, X_{d,p}$  are capable of causing. The assumptions about correlations between local effects  $\tau_d$  are identical to those for the models already discussed. More precisely, either we consider that these effects are mutually independent, which is simpler but perhaps sometimes inconsistent with the reality in the field, or we introduce spatial correlations, for example by using expressions 12.3 or 12.4. In both cases, it is a generalised linear mixed model.

## 12.2 Forming the "small area" estimator

Defining the model to be used is only a first step in the process. At this stage, we still only see the benefit of the model qualitatively, which reduces the scale of the problem by considerably simplifying reality. Indeed, it is much easier to make estimates in an environment where all the relevant information is assumed to be explained by a few well-known variables and by a few parameters rather than working in an undefined system that would consequently depend on an infinite number of uncontrolled components..., as assumed by classical sampling theory!

The next step is to choose the estimation strategy - we should also now talk about prediction since the parameter of interest has become a random variable following modelling.

### 12.2.1 BLUP estimation strategy: the standard individual model

In this section, we only consider linear models. In this context, several strategies for estimating/predicting the parameter of interest can be used but we now discuss what is probably the most common, the *Best Linear Unbiased Predictor* (BLUP) strategy. We consider the case where the parameter is mean  $\bar{Y}_d$ . Its predictor is generally a function of the data collected, *i.e.*  $Y_i$  where  $i$  describes the overall responding sample  $s$ . Above all, the statistician seeks a linear predictor of type  $\sum_{i \in s} a_i Y_i$  where  $a_i$  are real unbiased coefficients, *i.e.* its expected value equals that of  $\bar{Y}_d$ . Finally, the statistician seeks to minimise the mean square error which is the expected value of the square of the difference between the predictor and the value  $\bar{Y}_d$  it has to predict. The solution to this mathematical problem is the BLUP estimator (or predictor), also called the Henderson estimator in the literature. We will designate it  $\tilde{Y}_d^H$ .

In the specific case of the standard individual linear mixed model (see section 12.1.2), when the sampling fraction is negligible, we confirm that the BLUP estimator is written:

$$\tilde{Y}_d^H = \gamma_d [\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T \tilde{\beta}] + (1 - \gamma_d) \bar{X}_d^T \tilde{\beta} \quad (12.11)$$

All vectors are column vectors, the transposed vector being identified by exponent  $T$ . By designating  $D$  the total number of areas of interest, by designating  $X_i = (X_{i,1}, X_{i,2}, \dots, X_{i,p})^T$  the vector of auxiliary variables,  $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T$  the vector of the model parameters associated with these variables,  $\bar{x}_d = \frac{1}{n_d} \sum_{i \in s \cap d} X_i$ ,  $\bar{y}_d = \frac{1}{n_d} \sum_{i \in s \cap d} Y_i$  and  $\bar{X}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} X_i$ , we have:

$$\gamma_d = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \frac{\sigma_e^2}{n_d}} \quad (12.12)$$

$$\tilde{\beta} = \left( \sum_{d=1}^D \left( \sum_{i \in s \cap d} X_i X_i^T - \gamma_d n_d \bar{x}_d \bar{x}_d^T \right) \right)^{-1} \cdot \left( \sum_{d=1}^D \left( \sum_{i \in s \cap d} X_i Y_i - \gamma_d n_d \bar{x}_d \bar{y}_d \right) \right). \quad (12.13)$$

The coefficients vector  $\tilde{\beta}$  does not have a familiar expression here, but we can confirm that it is the classical and well-known estimator called "generalised least squares", frequently encountered in the linear regression model theory. It optimally estimates the vector of unknown parameters  $\beta$  of the model.

It is important to note that we need to know the true means by area  $\bar{X}_d$ . In practice, this means that individual variables  $X_i$  are available in a certain comprehensive file covering the scope of the survey (which does not mean that these individual values are accessible to the statistician responsible for the estimate, who perhaps only has  $\bar{X}_d$ ). However, this file may not be the sampling database and the  $X_i$  information used to calculate  $\tilde{\beta}$  may come from the survey collection file, in exactly the same way as  $Y_i$ . In this case, which is common in practice, it should be ensured that variable  $X$  derives from the same concepts in both sources (comprehensive file and collection file). For example, to calculate  $\tilde{\beta}$  from a labour force survey where  $X$  represents the employment status collected in the survey and to form  $\tilde{Y}_d^H$  using  $\bar{X}_d$  representing the employment status declared in the census would be very hazardous.

Formally, Henderson's estimator consist in two elements that are combined using real coefficient  $\gamma_d$ . The first element - located in the square brackets of Equation 12.11 - is a circumstance estimator that is a little complicated to interpret but that has the same statistical performance as  $\bar{y}_d$ , the estimator constructed from sub-sample  $s \cap d$ : it has a sampling variance as a decreasing function of  $n_d$ , so *a priori* large. Because this characteristic is associated with the direct estimators, and because at the same time the presence of coefficient  $\tilde{\beta}$  - formed from the complete sample - does not allow it to be qualified strictly as a direct estimator, we will talk about a pseudo-direct estimator. The second element is an estimator constructed by multiplying regression coefficient  $\tilde{\beta}$  by the true mean of auxiliary variable  $\bar{X}_d$ , which intuitively should give a value close to the true mean of the variable of interest if the model is appropriate. This estimator  $\bar{X}_d^T \tilde{\beta}$  is called a *synthetic estimator*. Its statistical properties are totally dependent on those of  $\tilde{\beta}$  since mean  $\bar{X}_d$  is not random. We can see for ourselves that  $\tilde{\beta}$  is made up of terms involving the entire responding sample  $s$  and not just the  $s \cap d$  part. By its nature, this makes it very stable, in other words weakly dependent on responding sample  $s$ . If we consider only the sampling uncertainty, we can therefore say that the synthetic component offers a low sampling variance. The flipside to this stability is the existence of a sampling bias, which may be numerically strong if the model is inappropriate.

Coefficient  $\gamma_d$ , which is always between 0 and 1, is a remarkable coefficient because it optimally weights (remember that it minimises the MSE) the two separate components, which have completely opposed behaviours in terms of both bias and sampling variance. In this, we say that  $\tilde{Y}_d^H$  is a *composite estimator* (or *mixed estimator*). The BLUP strategy therefore leads to an expression of  $\gamma_d$  that gives priority to the most efficient of the two components. We'll take the case where  $\sigma_\tau^2$  is small, which corresponds to small local effects  $\tau_d$ , *i.e.* to an efficient model, since it carries the true explanatory character on controlled auxiliary variables  $X_i$  and not on the "catch-all" residual term  $\tau_d$ . Under such circumstances, we tend to trust the model and build the final estimator based as much as possible on the model, *i.e.* the synthetic estimator. This is actually what happens since  $\gamma_d$  is small. Now let's take the case where the responding sample size  $n_d$  is large. Such a context gives confidence in the pseudo-direct estimator, which doesn't (or scarcely) uses the model and therefore by construction that is unlikely to be hindered by the model lacking relevance (the pseudo-direct estimator has weak bias, and in this case low variance since  $n_d$  is large). This is what we conclude since  $\gamma_d$  is large, close to 1.

We add that, with this theory, we can easily predict each local effect  $\tau_d$ . After simple but nonetheless tedious calculations, we get:

$$\tilde{\tau}_d = \gamma_d (\bar{y}_d - \bar{x}_d \tilde{\beta}) \quad (12.14)$$

which allows the Henderson estimator to be written in a more intuitive form:

$$\tilde{Y}_d^H = \bar{X}_d^T \tilde{\beta} + \tilde{\tau}_d. \quad (12.15)$$

There is still one step to be completed to reach the operational stage. Indeed, the BLUP estimator  $\tilde{Y}_d^H$  has a complex expression that at this stage depends on certain components of the vector of the parameters for model  $\delta$  introduced at 12.1.2. Indeed, applying the BLUP strategy made it possible to produce estimators  $\tilde{\beta}$  of  $\beta$  that have reduced the scale of the problem which means the vector of initial parameters  $\delta$  is now limited to variance components, *i.e.* two real values  $\sigma_\tau^2$  and  $\sigma_e^2$ . They will be summarised by vector  $\Sigma = (\sigma_\tau^2, \sigma_e^2)$ . In fact, what we call - commonly but incorrectly - estimator  $\tilde{Y}_d^H$  is not one since this expression is not calculable and we should therefore strictly designate it  $\tilde{Y}_d^H(\Sigma)$  and talk about a "pseudo estimator". As components of  $\Sigma$  are unknown, they will have to be estimated using the data collected. Once parameter  $\Sigma$  has been estimated by

$\hat{\Sigma}$ , we will substitute  $\hat{\Sigma}$  with  $\Sigma$  in  $\tilde{Y}_d^H(\Sigma)$  to arrive at a new expression, i.e.  $\tilde{Y}_d^H(\hat{\Sigma})$ , which this time deserves the name of estimator since it can be calculated. We give the estimator/predictor obtained in this way the name *Empirical Best Linear Unbiased Predictor* (EBLUP).

We frequently estimate  $\Sigma$  by the maximum likelihood method. We also have a variant called restricted maximum likelihood, which can be recommended as it reduces the bias of estimators when sample sizes are modest. Nonetheless this approach imposes an additional assumption on the distribution of random variables  $\tau_d$  and  $e_i$ , which is almost systematically considered as variables following a Gauss distribution. There are no analytical expressions giving  $\hat{\sigma}_\tau^2$  and  $\hat{\sigma}_e^2$ , but numerical algorithms are able to produce estimates matching the theory. Based on these estimates, we get

$$\hat{\gamma}_d = \frac{\hat{\sigma}_\tau^2}{\hat{\sigma}_\tau^2 + \frac{\hat{\sigma}_e^2}{n_d}} \text{ then:}$$

$$\hat{\beta} = \left( \sum_{d=1}^D \left( \sum_{i \in s \cap d} X_i X_i^T - \hat{\gamma}_d n_d \bar{x}_d \bar{x}_d^T \right) \right)^{-1} \left( \sum_{d=1}^D \left( \sum_{i \in s \cap d} X_i Y_i - \hat{\gamma}_d n_d \bar{x}_d \bar{y}_d \right) \right) \quad (12.16)$$

and finally the EBLUP estimator:

$$\hat{Y}_d^H = \hat{\gamma}_d [\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T \hat{\beta}] + (1 - \hat{\gamma}_d) \bar{X}_d^T \hat{\beta}. \quad (12.17)$$

Note that we can avoid any assumption relating to the distribution of  $Y_i$  by using a "method of moments", but conversely it proves theoretically less effective if the distribution of random variables  $\tau_d$  and  $e_i$  is indeed Gaussian.

### 12.2.2 Application to the individual linear model with spatial correlation

The BLUP strategy, with its natural extension EBLUP, is applied in exactly the same way when spatial correlations between local effects are introduced. The difference with the standard linear model lies solely in the mathematical expressions of the various estimators, which are obviously much more complicated, but the principles do not change. Detailing the formal expression of the Henderson estimator in the presence of spatial correlations can only reasonably be done by using matrix-based notation, which is burdensome and has no added educational value.

The BLUP (or EBLUP) estimator remains a combination of a direct estimator and a synthetic estimator, with an optimal weighting calculated taking the context into account, depending on the confidence that can be given to the model and the responding sample size  $n_d$ . The coefficient  $\sigma_\tau^2$  introduced into Equation 12.3 retains an essential role, but the calculations must now be done also taking the additional coefficient  $\rho$  into account, which adjusts the intensity of spatial correlation. The model parameter to be estimated is therefore  $\Sigma = (\rho, \sigma_\tau^2, \sigma_e^2)$ .

The algorithms for calculating maximum likelihood (restricted, if applicable) adapt to the introduction of an additional parameter, and they produce an estimate of  $\rho$ ,  $\sigma_\tau^2$  and  $\sigma_e^2$ . The complexity of the variances-covariances structure does not seem to allow methods for estimating  $\Sigma$  other than maximum likelihood or restricted maximum likelihood.

### 12.2.3 Application to the Fay and Herriot model

The Fay and Herriot model is very important as, in practice, it is widely used. In many cases, it fits well and produces satisfactory estimates, which are preferable to direct estimates. Although it involves a higher degree of aggregation than in the previous models, the BLUP strategy is also implemented within this model. In the expression of Henderson's optimum estimator with the standard model, the  $\sigma_e^2$  terms have obviously disappeared but on the other hand we find the values

of true sampling variances by area  $\psi_d$ . It is important to note that in the standard theory, true sampling variances are assumed to be known. This is obviously not the case in reality, and *in fine* we have to replace theoretical expressions  $\psi_d$  by estimators  $\hat{\psi}_d$  obtained by applying traditional methods to calculate sampling variance. At this stage, it is recommended to finish by smoothing the  $\hat{\psi}_d$  values. This operation protects against including abnormally weak or abnormally strong  $\hat{\psi}_d$  estimates, thus avoiding a highly adverse impact on the quality of final estimates by area. We end up at

$$\tilde{Y}_d^H = \gamma_d \hat{Y}_d + (1 - \gamma_d) \bar{X}_d^T \tilde{\beta} \quad (12.18)$$

with

$$\begin{aligned} \gamma_d &= \frac{\sigma_\tau^2}{\sigma_\tau^2 + \hat{\psi}_d} \\ \tilde{\beta} &= \left[ \sum_{d=1}^D \frac{\bar{X}_d \bar{X}_d^T}{\sigma_\tau^2 + \hat{\psi}_d} \right]^{-1} \cdot \left[ \sum_{d=1}^D \frac{\bar{X}_d \hat{Y}_d}{\sigma_\tau^2 + \hat{\psi}_d} \right]. \end{aligned} \quad (12.19)$$

Estimator  $\tilde{Y}_d^H$  retains a composite form and the BLUP strategy produces the ideal  $\gamma_d$  weighting, shared between direct estimate  $\hat{Y}_d$ , which is independent of the model but unstable, and synthetic estimate  $\bar{X}_d^T \tilde{\beta}$  that is totally dependent on the model but, conversely, fairly insensitive to the composition of the responding sample. In rare cases, when estimating a proportion, estimate  $\tilde{Y}_d^H$  may fall outside the interval  $[0, 1]$ . In this case, the initial model must be adapted.

If spatial correlations are introduced, the above expressions change as a result - by becoming considerably more complicated - but none of the main principles are altered. In all cases, with or without spatial correlations, the software is able to produce estimator  $\sigma_\tau^2$  using maximum likelihood (restricted if necessary), from which we immediately deduce  $\hat{\gamma}_d$  and  $\hat{\beta}$ , then the final EBLUP estimator  $\hat{Y}_d^H$ . Note that in the absence of spatial correlation, there are other methods for estimating parameter  $\sigma_\tau^2$  besides maximum likelihood.

#### 12.2.4 Strategy for non-linear models

The world of non-linear models is technically much more complicated than that of linear models. In particular, the BLUP strategy is not directly suited to this context because there is no satisfactory mathematical solution. Nonetheless, it remains a basic technique and that is why one way to deal with non-linear models, such as the logistic model or the Poisson model, is to replace them with approximate models that have a linear structure. What an approximate model is refers to a complicated but nonetheless operational theory. The BLUP strategy is applied by starting from the approximate linear model.

The initial model may or may not use spatial correlations. The developments presented in the preceding sections are then applied to the approximate linear model.

However, the most compelling approach is to use a strategy better suited to this non-linear context, such as the *Empirical Bayes* strategy, which produces optimal estimates, or the *Hierarchical Bayes* strategy, which corresponds to the classical Bayesian approach.

### 12.3 The quality of estimators

The model approach will obviously result in making the estimate dependent on the choice of the model and will therefore raise the question of the relevance of the model used. Indeed, simplification has a cost in terms of quality and one may question how correctly this model represents reality.

### What are we talking about?

In terms of assessing the quality of small area estimations, it is more necessary than ever to specify the concept of quality. Indeed, the context suffers from a very specific complication due to the coexistence of different kinds of uncertainties, on the one hand, the sampling uncertainty that decides on the composition of the sample, and on the other hand, the uncertainty of the model which handles the variable of interest as a random variable. Quality can be assessed with or without taking the model uncertainty into account.

If the modelling does not include any uncertainty, one is dealing with the survey statistician's classical approach placed in finite population and handling with deterministic individual variables. From this viewpoint, the situation is extremely simple because all the "small area" estimators presented up to now are biased. This is the natural consequence of the failure to take sampling weightings into account (when sampling is not at equal probabilities in all cases), or only partial consideration of these weightings. For example, in the individual standard linear model, the weighting reflecting sampling is always missing. In the Fay and Herriot model, it is certainly found in the direct component  $\hat{Y}_d$  but not at all in the synthetic part  $\bar{X}_d^T \hat{\beta}$ . On the other hand, the model provides a decisive advantage in terms of sampling variance because the  $\beta$  parameters that are estimated use the entire responding sample and hence have a weak sampling variance. The estimated local random effect  $\hat{\tau}_d$  is unstable but if the model is well suited, it will be numerically small and its variance will therefore have limited influence. The Henderson estimator should ultimately have limited sampling variance and *a priori* less than for the direct estimator if the model has good explanatory power.

When taking model uncertainty into account, if the model is linear by construction, the BLUP estimator is unbiased. Switching to EBLUP only incurs negligible bias. If the model is not linear, the context is much more complicated, but modest bias is expected.

#### 12.3.1 An iterative process

Quality assessment can be designed using a cyclical mechanism (Figure 12.1).

Having, on the one hand, certain selection criteria for explanatory variables, and on the other, having a set of auxiliary variables  $X$  potentially explanatory for  $Y$ , we adjust a model. At this stage, we have statistical tools to assess the quality of this adjustment. Combined with a prediction strategy, this model produces a theoretical estimator  $\tilde{Y}_D$ . This estimator depends on parameters that contribute to defining the model (at least parameter  $\sigma_\tau^2$ ,  $\sigma_e^2$  if applicable and  $\rho$  if there is a spatial correlation). These parameters are estimated using an *ad hoc* method. At the end of the cycle, we assess the quality of the final predictor (bias, MSE; see sections 12.3.3 and 12.3.4). If it is not acceptable, a new cycle is initiated by re-examining the relevance of the model, or even the relevance of the prediction strategy or the estimation of the model parameters. Quality assessment also involves checking the relevance of the model's distribution assumptions, if necessary. This is why we will confirm the Gaussian nature of estimated local effects  $\hat{\tau}_d$  when a maximum likelihood (whether restricted or not) technique has been used.

#### 12.3.2 The problem of bias

Survey statisticians are sometimes reluctant to use a model-dependent estimator (although it is essential to handle non-response). Their main fear is of substantial bias if we confine ourselves to sampling uncertainty. This risk is inevitable since the model simplifies, and therefore distorts, reality. The important thing is not to escape the bias but to obtain limited bias that is more than offset by the gain in terms of variance. Unless we are working on artificial populations, calculations of bias due to sampling cannot be done, but two simple tools can be used to assess the situation, however without providing evidence.

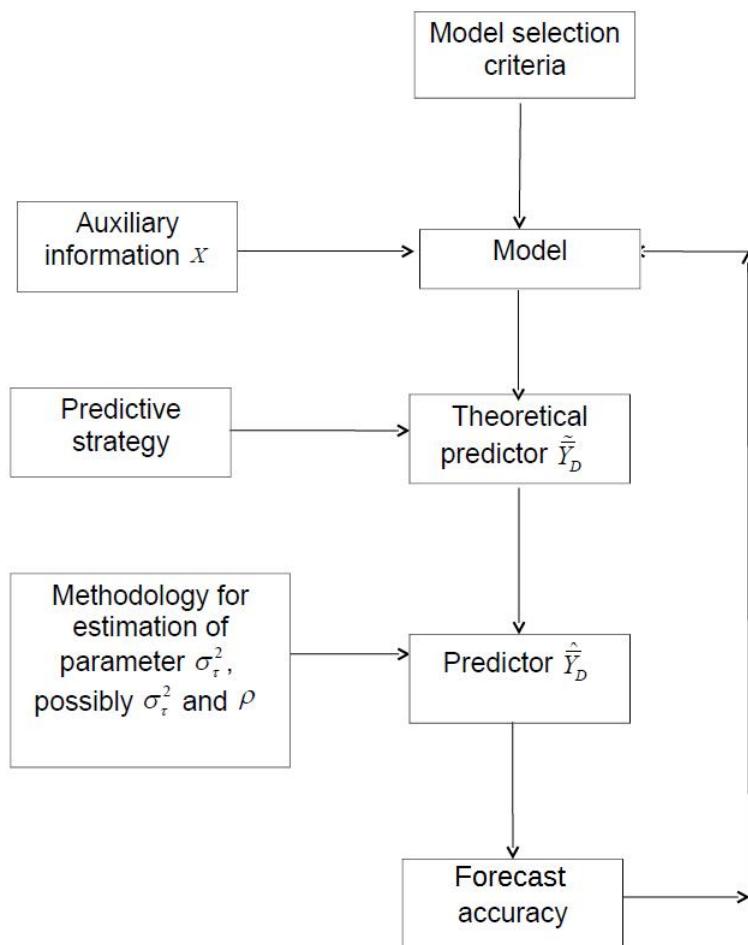


Figure 12.1 – Diagram of the iterative process for assessing the quality of estimators

The first tool is purely graphical and consist in constructing a point cloud where each point represents one of the  $D$  areas handled. One of the axes plots the direct estimate (therefore obtained without a model), the other axis plots the "small area" estimate (therefore from a model). If the resulting point cloud is not symmetrical around straight line  $y = x$  (first bisector), there is a strong suspicion of bias due to sampling. Nonetheless, there is no inevitability (remember the situation, obviously idealised, of a model reflecting a reality in which all means by area are equal). The converse situation is more convincing in the sense that if the point cloud is symmetrical, there will probably be no significant bias due to sampling. Most often, in practice, we see that scatter points form an angle with the first bisector that, when projected onto the axis representing the "small area" estimate, is smaller than the projection onto the axis representing the direct estimate. This phenomenon is called *shrinkage*, and it is therefore rather an indication of bias due to sampling. It reflects a form of *essentially* excessive concentration of estimates. It arises mechanically from the simplifying model, which has a normalising effect and therefore more or less tends to standardise estimates by area. We stress that this graphical approach offers no evidence but only creates suspicions. In practice, because it cannot accurately reflect reality, any model inevitably creates a theoretical bias due to sampling and the possible symmetry of the point cloud only indicates the probable weakness of this bias.

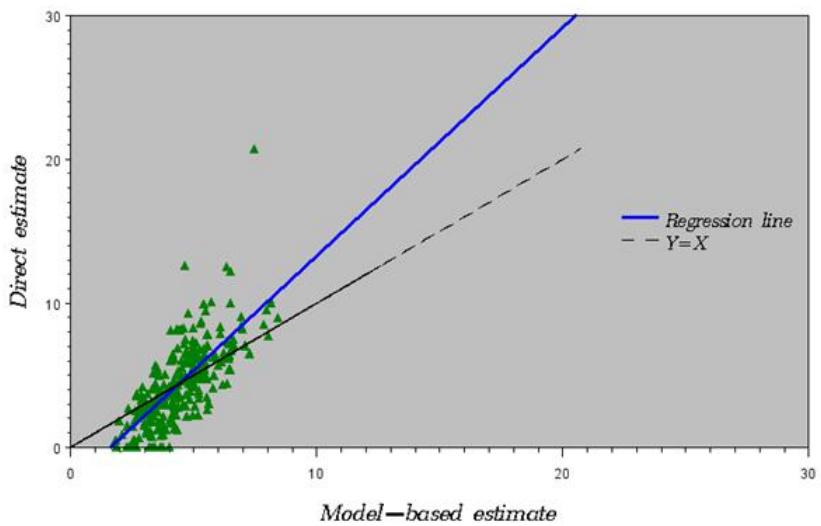


Figure 12.2 – Example of relationship between direct estimates and small area estimates

The second technique is even simpler and more intuitive as it involves summarising estimates of totals  $\hat{T}_d^H$  obtained on  $D$  small areas and comparing the result with the direct estimate of the total  $\hat{T}$  covering the whole population that results from classical sampling theory. In fact, the latter is by construction unbiased (the uncertainty here is exclusively sampling uncertainty). If there is a bias due to the model and this bias is somewhat systematic, a discrepancy will be seen between the two values. On the other hand, a bias without a systematic component cannot be detected since compensation may occur during summation.

It is customary to use the above-mentioned difference in order to increase quality. Indeed, if  $\hat{T}_d^H$  is the "small area" estimator of true total  $T_d$  in area  $d$ , if  $\hat{T}$  is the unbiased direct estimator from the overall responding sample  $s$  representing the whole population  $U$ , we very often adopt the

following final estimate:

$$\hat{\hat{T}}_d^H = \hat{T}_d^H \frac{\hat{T}}{\sum_{d=1}^D \hat{T}_d^H} \quad (12.20)$$

which makes it possible to calibrate the estimate of the total in  $U$  on  $\hat{T}$ . This operation is called *benchmarking* and helps to limit the bias of  $\hat{T}_d^H$  while ensuring consistent distribution.

Furthermore, it is always interesting to map the mean estimates by area  $\hat{Y}_d^H$ , which provides a visual check on the consistency of the estimation system as a whole. Normally, two areas with similar and related characteristics on a map should correspond to two similar estimated means  $\hat{Y}_d^H$  (in practice, the colours representing their respective values should be in the same range).

### 12.3.3 Mean square error

In an environment where bias is possible, probable, or even inevitable, the correct error concept is that of mean square error (MSE). This indicator designates the expected value of the square of the difference between the estimator and the parameter. Taking into account both sampling uncertainty and model uncertainty, the theoretical framework offered by the model makes it possible to obtain the expression for MSE and then estimate it without or almost without bias. The expression for MSE and its estimation are very complicated, even with the linear model, and the calculation is therefore entrusted to a software package. Nonetheless, without spatial correlation, we can confirm, if there is a large number of areas  $D$ , that the numerically more important term in estimating the MSE of  $\hat{Y}_d^H$  is  $\hat{\gamma}_d \hat{\psi}_d$  for the Fay and Herriot model, and  $\hat{\gamma}_d \frac{\hat{\sigma}_e^2}{n_d}$  for the standard individual linear model. Regarding all these error calculations, the results obtained essentially assume that the model is specified in a way that perfectly matches reality (the model can be qualified as "exact"). This is certainly not strictly true! Introducing spatial correlation obviously creates an additional technical difficulty, but general theory makes it possible to succeed, which does not mean that currently available information technology tools are able to use it. Note that under certain particularly favourable circumstances, we have an external source that can provide the true value of the parameter (e.g. after a census). This makes it possible to assess the estimation error made directly.

## 12.4 Implementation with R

■ **Example 12.1 — Dissemination of census on squares.** In 2021, the European Union Statistical Office EUROSTAT wishes to produce statistics (gender, age range, activity, etc.) covering the entire population of each Member State in one-by-one kilometre squares. In addition, in France, INSEE aims to disseminate data from the Population Census using squares whose side could measure some hundred meters. Since 2004, the French census has been carried out by sampling in the municipalities with more than 10,000 inhabitants<sup>2</sup>. Therefore, the targeted area contains too few observations to obtain good direct estimators of the parameters of interest. This is why *small area* estimation is an appropriate statistical technique for using this type of data.

Introducing a spatial correlation in this context makes it possible to reflect the phenomenon of natural continuity of the socio-demographic characteristics of individuals populating geographically contiguous areas. Indeed, moving from any square to neighbouring squares, one cannot reasonably assume there is independence between the behaviours of the statistical units – households or individuals – which formed it. ■

2. The census is complete (exhaustive) in municipalities of fewer than 10,000 inhabitants

The *sae* package in R is used to calculate small area estimates at "area" and "individual" levels, in the case of models respectively not taking and taking into account spatial autocorrelation. This package, which was implemented by Molina and Marhuenda, was described in *The R Journal* (Molina et al. 2015).

The main functions that have been used to handle data from the French census, using a model developed at area level, are from the *sae* package. They are `eblupFH()`, `eblupSFH()`, `mseFH()` and `mseSFH()`.

To produce estimates using a model at individual level, we used functions `eblupBHF()` and `pbmseBHF()` from the *sae* package, as well as function `corrHLfit()` from the *spaMM* package.

### Area level modelling: `eblupFH()` and `mseFH()` base functions

The first estimates are based on the Fay and Herriot model without spatial autocorrelation. Function `eblupFH()` provides as output:

- i) Fay and Herriot estimates for each area;
- ii) an estimate of variance  $\sigma_\tau^2$  of the random effect specific to the areas.

In addition, function `mseFH()` produces the calculation of mean square errors associated with each estimate (see section 12.3.3).

The arguments for these functions are the same. The standard syntax is as follows:

---

```
mod_FH <- eblupFH(formula = Y ~ X1+...+Xp, vardir = varech, method = ,
  maxiter = m, precision = e, B = 0, data = )
```

---

First, the `formula` parameter specifies the variable of interest  $Y$  as well as the explanatory variables selected  $X_1, \dots, X_p$ . The numerical values of all these variables must be contained in a table that associates one line with each area, specified in the argument `data`. The  $\hat{\gamma}_d$  parameters involved in the Henderson estimators are calculated using (estimated) sampling variances per area  $\hat{\psi}_d$ , which are available in a variable specified in the argument `vardir`.

The estimate of the sole variance parameter of the model is obtained using an *ad hoc* technique. In practice, we use an iterative method that should converge towards value  $\sigma_\tau^2$ . The `maxiter` and `precision` parameters are technical parameters (defined either by the user or by default) that govern this iterative process. The algorithm calculates an estimate of  $\sigma_\tau^2$  at each stage of iteration. The role of the `precision` parameter is as follows. As soon as the difference between two consecutive values is less than this ( $e$  in our example), the algorithm stops. Otherwise, as long as the maximum number of iterations `maxiter` is not reached, iterations continue. The output indicates whether or not the algorithm converges. The method must also be specified. We can choose from three methods, including the maximum likelihood method and the restricted maximum likelihood method (respectively `method = "ML"` and `method = "REML"`). The third method (`method = "FH"`) is a "method of moments".

The reader's attention is drawn to the need not to add dummy variables identifying the areas to the regressors. Indeed, as the constant is already a part of one of the standard regressors, this practice would lead to create a non-invertible matrix. The following command therefore leads to a failure:

---

```
mod_FH <- eblupFH(formula = Y ~ X1+...+Xp+as.factor(Carreau) , vardir =
  varech, method = , maxiter = m, precision = e, B = 0, data = )
```

---

### Spatial correlation at area level: the `eblupSFH()` and `mseSFH()` functions

The software estimates a SAR model (see section 12.1.3) of the type

$$\tau = \rho \cdot A \cdot \tau + u \quad (12.21)$$

The parameters of these functions are the same as for the previous functions, except that the proximity matrix **A** (coefficient matrix  $\alpha_{i,j}$ ; see section 12.1.3) must also be specified. The R command is as follows:

---

```
mod_SFH <- eblupSFH(formula = Y ~ X1+...+Xp, vardir = varech, method = ,
maxiter = m, precision = e, proxmat = A, B = 0, data = )
```

---

The proximity matrix is described by the `proxmat` parameter. It has standardised lines in that the sum of the elements of each line always equals 1.

A similar iterative process to that for use without spatial correlation makes it possible to calculate:

- i) Fay and Herriot estimates for each area;
- ii) an estimate of the variance of the random effect specific to areas;
- iii) an estimate of the spatial autocorrelation parameter  $\rho$ .

### Modelling at individual level, without spatial correlation: functions `eblupBHF()` and `pbmseBHF()`

Using individual level modelling, the `eblupBHF()` function from the *sae* package allows direct estimates and "small area" estimates to be calculated without spatial correlation. The syntax is as follows:

---

```
mod_BHF <- eblupBHF(formula = Y ~ X1+...+Xp, dom = ,
meanxpop = , popnsize = Popn, data = adr_est)
```

---

It uses the following configuration — `formula` for the formal expression of the model, `dom` to designate the variable identifying the areas, `popnsize` for the size of the population  $N_d$  in each area, `meanxpop` for the means of explanatory variables  $\bar{X}_d$  calculated in the whole population of the area. The `data` parameter designates the data table.

The `pbmseBHF()` function estimates the errors (MSE) of "small area" estimators using a *bootstrap* technique. The parameters of this function are the same as for the previous function, to which the number of re-samplings of the *bootstrap* defined by the `B` parameter is added (`B=1000` for example).

---

```
mse_BHF <- pbmseBHF(formula = Y ~ X1+...+Xp, dom = ,
meanxpop = , popnsize = , B = 1000, data = )
```

---

### Taking spatial correlation into account in the individual model

The *spaMM* package can be used to take spatial correlation into account. It can manage several types of models, in particular the Poisson model (see section 12.1.4). Function `corrHLfit()` handles the Poisson model at the individual level with spatial correlation.

---

```
library(spaMM)
mod_spa <- corrHLfit(formula = Y ~ X1+...+Xp+Matern(1|x+y),
HLmethod = "REML", family = "poisson", ranFix = list(nu=0.5), data = )
```

---

For configuring this function, `formula` designates the formal expression of the model. The *Matern(1|x+y)* component, which is specific to the function used, takes into account the coordinates *x* and *y* of the areas (here, the centres of the squares), which should therefore be contained in the

data table, in order to calculate the distances used in the spatial correlation function. Furthermore, `HLmethod` specifies the method for estimating variance and spatial correlation parameters (here, restricted maximum likelihood), `family` chooses the distribution of the variable of interest (here, a Poisson distribution). The functional form of the spatial correlation can be selected from a configured family of complicated functions called Matérn functions. Parameter `ranFix` specifies the configuration of this family of functions. If we indicate `list(nu=0.5)`, we get the exponential form of Equation 12.3, which is the expression traditionally used - except that the estimated parameter is  $\frac{1}{\rho}$  and not directly  $\rho$ . The `data` parameter designates the data table.

As output, we obtain, among other things, the estimated coefficients of the model, including the coefficient  $\rho$  involved in the exponential spatial correlation (in fact, the reverse if we refer to expression 12.3), optimal predictions  $\hat{\tau}_d$  of random local effects, and the estimated variance of random effect  $\hat{\sigma}_\tau^2$ .

## Conclusion

The small area estimate is based on the use of stochastic models. It is the counterpart to a certain shortage of information collected using the sample dividing the area when it is small. In order to limit inaccuracy, unsurprisingly, we have to make assumptions covering the entire population and compensating for the lack of information obtained at local level. Models explicitly involve local geographical effects, the interpretation of which is delicate, in that we can always consider it as a last resort to conceal inadequate consideration of explanatory fixed effects of the phenomenon studied. Basically, the first question is knowing to what extent there is a purely geographical effect. Moreover, these models, whatever they are, always create bias in relation to sampling uncertainty. The main aim is to limit its extent, rather than measure sampling variance, which becomes a secondary aim for the sampling statistician. Of course, we have statistical tools to assess the quality of the adjustment to a model, but this does not guarantee the selected model is suitable for the particular situation of a given area, which may be very specific without the statistician being aware of it. There is no reliable estimate for sampling bias, and we currently have only a few qualitative tools available, which are convincing to varying degrees and which only lead to an assessment of an overall situation. In general, the theory of linear models (*Linear Mixed Models* or LMM) is much simpler than that traditionally used (*Generalised Linear Mixed Models* or GLMM) for non-linear models, which are still really difficult to access. The presence of spatial correlation always complicates the context and then raises the question of the availability of the computer code to make the estimates. The development of R is promising and, in the future, we should move towards extending the range of models accepting spatial correlation.

**References - Chapter 12**

- Battese, George E, Rachel M Harter, and Wayne A Fuller (1988). « An error-components model for prediction of county crop areas using survey and satellite data ». *Journal of the American Statistical Association* 83.401, pp. 28–36.
- Chandra, Hukum, Ray Chambers, and Nicola Salvati (2012). « Small area estimation of proportions in business surveys ». *Journal of Statistical Computation and Simulation* 82.6, pp. 783–795.
- Coelho, Pedro S and Luis N Pereira (2011). « A spatial unit level model for small area estimation ». *REVSTAT-Statistical Journal* 9.2, pp. 155–180.
- Fay III, Robert E and Roger A Herriot (1979). « Estimates of income for small places: an application of James-Stein procedures to census data ». *Journal of the American Statistical Association* 74.366a, pp. 269–277.
- Molina, Isabel and Yolanda Marhuenda (2015). « sae: An R package for small area estimation ». *R Journal, in print*.
- Pratesi, Monica and Nicola Salvati (2008). « Small area estimation: the EBLUP estimator based on spatially correlated random area effects ». *Statistical methods and applications* 17.1, pp. 113–141.
- Rao, John NK (2015). *Small-Area Estimation*. Wiley Online Library.

# IV

## Part 4: Extensions

13	Graph partitioning and analysis .....	327
14	Confidentiality of spatial data .....	349
	Index .....	375



# 13. Graph partitioning and analysis

PASCAL EUSEBIO, JEAN MICHEL FLOCH, DAVID LEVY  
INSEE

---

<b>13.1</b>	<b>Graphs and geographical analysis of city networks</b>	<b>328</b>
13.1.1	Small World .....	328
13.1.2	Free scale networks .....	331
<b>13.2</b>	<b>Graph partitioning methods</b>	<b>333</b>
13.2.1	Concepts in graph theory .....	333
13.2.2	Partitioning methods .....	337

---

## Abstract

To analyse the network of cities in this study, we had to move away from the methods usually used at INSEE and resort to graph-based representations. While these techniques are still not widely used in public statistics, the problem raised is fairly standard — partitioning the population into sub-populations. This consists in identifying homogeneous (with low intra-class heterogeneity) and quite differentiated (high inter-class heterogeneity) sub-populations. Using graphs, we will see that we often look for partitions that maintain a large number of intra-zone flows and little flow between them. Algorithmic solutions are based on “agglomerative” or “division” methods, depending on the case, which can be likened to the bottom-up or top-down methods we are familiar with in data analysis. They use the concept of modularity, based on comparing the graph studied with a random graph.

This chapter is not intended as a comprehensive review of all graph theory methods, which have undergone major changes since they first emerged in the 1930s. Such methods have been developed in a very wide range of fields (geography, social media analysis, biology, IT). The methods presented here are derived mainly from the world of physics (around the key modularity concept). However, one box provides some additional information on *block modelling* methods, and on how to take space into account in the networks.

## 13.1 Graphs and geographical analysis of city networks

Geographers have long taken an interest in analysing relations between territories. A great deal of work has been done on urban hierarchies. One of the most prominent examples is Christaller 2005's theory of central locations. The amount of data and processing tools available have long been a limit to flow analysis. The gravity models from Wilson's work have been a simple way to model interactions (Wilson 1974). The situation has changed considerably with specific developments in graph theory, derived from fields other than geography (sociology, as regards some intuitions, physics, IT). Two graph models proved of particular importance — small-world graphs and free scale graphs.

**Definition 13.1.1 — Graph.** A **graph** is a graphical representation of a set of vertices connected by edges.

An **edge** is a link between two separate items.

A **vertex** or **node** is an element connected by edges. The **degree** of a vertex is the number of vertices with which it is linked.

■ **Example 13.1** A graph of cities depicts cities (vertices) exchanging populations. Commuters are represented on edges, also referred to as links hereafter. ■

### 13.1.1 Small World

For many years, graph specialists were only interested in the random graphs still widely in use today. In the 1990s, various graph theorists offered models such as small-world and free scale. These models had a certain impact on geographical analysis. Small-world graphs were proposed by Watts and Strogatz in an article in the magazine *Nature* (Watts et al. 1998). Figure 13.1 shows the reproduction of the diagram proposed by the two authors to illustrate the construction of the small-world graph.

**Definition 13.1.2 — Random graph.** A graph with random edges distribution.

The idea of the small-world principle originates in Stanley Milgram's work. In his experiment, Milgram asked residents of the American Mid-West to send a letter to a West Coast recipient whom they did not know, using people around them as their intermediaries. Milgram was surprised to see that on average the chains leading to the recipient were made up of only 5.6 individuals. The experiment confirmed the theory Karinthy 1929 that everyone in the world is connected by a chain of no more than 5 links, which has become, in its popular version, the six degrees of separation. In plain language, only five people separate us from any other person in the world.

The starting graph is referred to as a k-regular graph.

**Definition 13.1.3 — K-regular graph.** A graph in which each vertex is connected with the same number of  $k$  vertices (Battiston et al. 2014). In other words, all vertices have the same  $k$  degree.

The authors wished to show, in a simple manner, how this regular graph could be turned into a random graph. At each stage, a link is randomly deleted with probability  $p$ , and a link added in the same manner. The process is described in detail in the founding article. Watts and Strogatz combined two measurements,  $L(p)$  and  $C(p)$ , to characterise a type of network.

$L(p)$  Refers to the average length of the shortest path between pairs of vertices when  $p$  varies.  $C(p)$  refers to the *clustering* coefficient, an illustration of which can be found in Figure 13.2. This coefficient is connected with the concept of transitivity in the graph, a concept known to sociologists since the 1970s. The idea of transitivity can be easily translated by the fact that our friends' friends are often our friends. High transitivity in the graph means that, from a topological point of view,

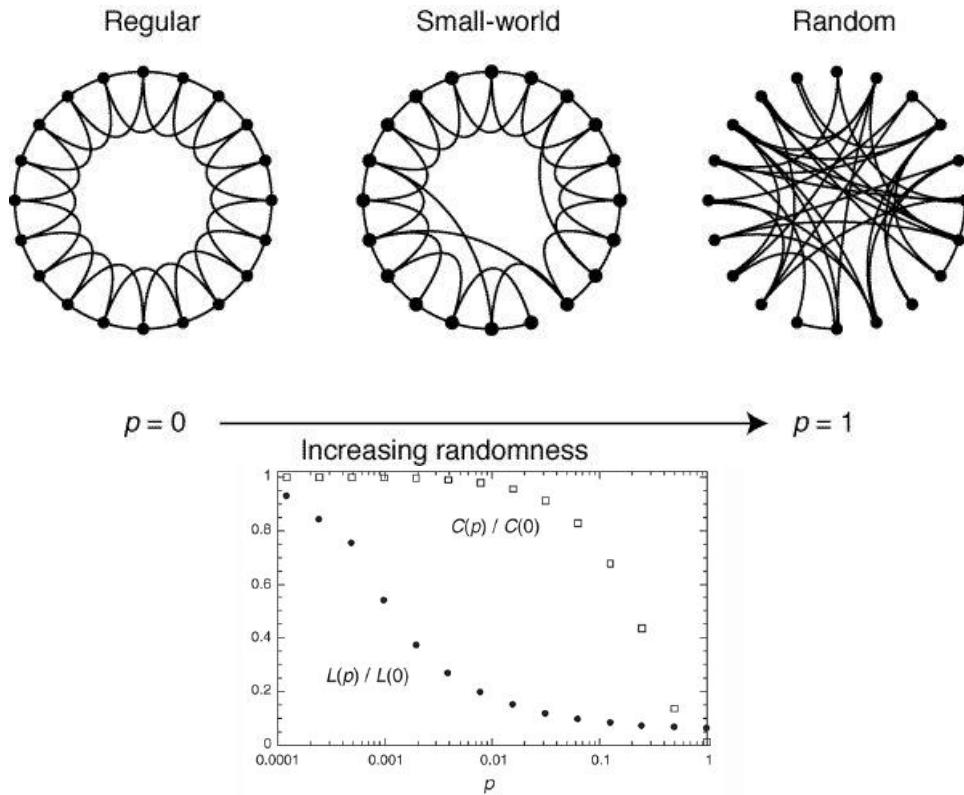


Figure 13.1 – Small-world networks

**Source:** Watts et al. 1998

there are many triangles. Strogatz and Watts suggested local (associated) coefficients at each node of the graph, and a global coefficient, which is the arithmetic average of the local coefficients.

**Definition 13.1.4 — The *clustering coefficient*.**

$$C(p) = \sum_{i=1}^n \frac{C_i}{n} \quad (13.1)$$

with

$$C_i = \frac{\text{number of triangles of which one of the three vertices is node } i}{\binom{k}{2}} \quad (13.2)$$

where  $k$  is the local coefficient or degree of the node and  $n$  the number of nodes in the graph.

■ **Example 13.2** With the graph shown in figure 13.2,

$$C_4 = \frac{5}{\binom{6}{2}} = 1/3$$

and the *clustering coefficient* of the network is:

$$C(p) = \sum_{i=1}^n \frac{C_i}{n} = 0.5208.$$

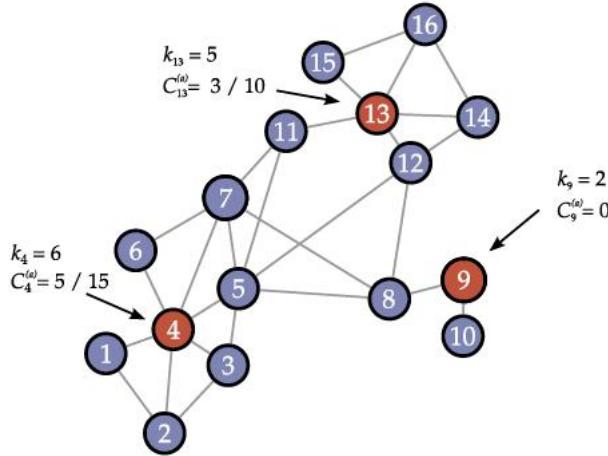


Figure 13.2 – The *clustering* coefficient

The values of  $C(p)$  and  $L(p)$  are standardised by the values  $C(0)$  and  $L(0)$  reflecting a regular graph. The two indicators develop very differently. The average distance between nodes decreases rapidly while that of the *clustering* coefficient (ratio of the number of triangles on the number of possible triplets) remains stable for a moment and decreases more rapidly. Watts and Strogatz deemed that for intermediate values of  $p$ , the networks remained fairly highly structured, like regular graphs, but with a low average length in paths, as in random graphs. They have called these small-world graphs, in a definition that remains largely qualitative — large number of vertices, number of existing links far from saturation, high degree of *clustering*, low average distance. More precise mathematical definitions have been proposed, but are also very technical and are beyond the scope of our study.

Small-world networks can be generated using the `sample_smallworld` function of the *igraph* package in R. In such a network, it is assumed that each vertex can be linked to any other.

### Application with R

---

```
# Necessary package
library(igraph)

# Graph generation with 100 nodes
g <- sample_smallworld(dim = 1, size = 100, nei = 5, p = 0.05)

# Graph representation
plot(g, vertex.size=4,vertex.label.dist=0.5,
      vertex.color="green",
      edge.arrow.size=0.5)


---


# graph coefficient calculation
## local coefficient
q=transitivity(g,type = "local")

## overall coefficient
transitivity(g,type = "average")
```

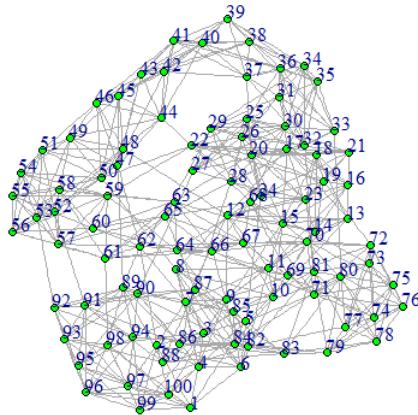


Figure 13.3 – Small-world graph

**Source :** Simulation from *igraph* package

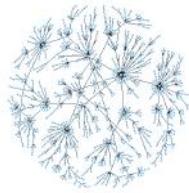
```
# which is well equal to the average of the local coefficients
mean(q)
```

### 13.1.2 Free scale networks

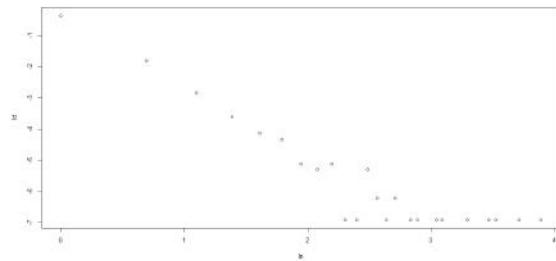
Another complex set of graphs can be found in free scale graphs. This type of modelling was initially proposed by Barabási et al. 1999. This type of graph can be generated under R with the *barabasi.game* function of package *igraph*, an illustration of which can be found in figure 13.4.

The logic behind the creation of this type of graph differs in particular from that of the small-world principle. These graphs show a particular distribution of degrees that is similar in type to that of the (Barabási et al. 1999) power law. Each new node will have a probability of binding to a node that is all the greater as the degree of this node higher. They are called invariant, because zooming in on any part of the graph does not change its shape. At each level of magnification, the network will contain a few nodes with many connections and a large number of nodes with very few connections. Thus, the network is said to be **free-scale** if, when  $k$  refers to the degree, and  $P(k)$  the frequency of vertices of degrees  $k$ , estimating the function  $P(k) = k^{-\gamma}$  shows a value of  $\gamma$  greater than 2. In the example shown in figure 13.4, the value of coefficient  $\gamma$  is 2.6.

The two models, briefly described here, do not exhaust the description of complex networks. In a book, Newman (author of several graph partitioning algorithms), Barabási (introducer of free scale graphs) and Watts (small-world graphs) show that complex graphs often combine characteristics of both types (Newman et al. 2011). This is very clear in the urban networks which we will now address. We often encounter communities of cities with strong interactions (small-world characteristics) while at the upper level, the links between communities are more a matter of invariance of scale. Numerous works have been carried out on city networks. Various works can be cited from Rozenblat et al. 2013 on air transport networks, on the combination of air and sea transport, or on geographical links between multinational companies. Figure 13.5 consists in a diagram showing the connections between small-world logic and free-scale logic.



(a) Example of Barabasi graph



(b) change in frequency of number of neighbours

Figure 13.4 – Example of free scale networks

**Source:** *Graphs simulated by the barabasi.game function in the igraph package*

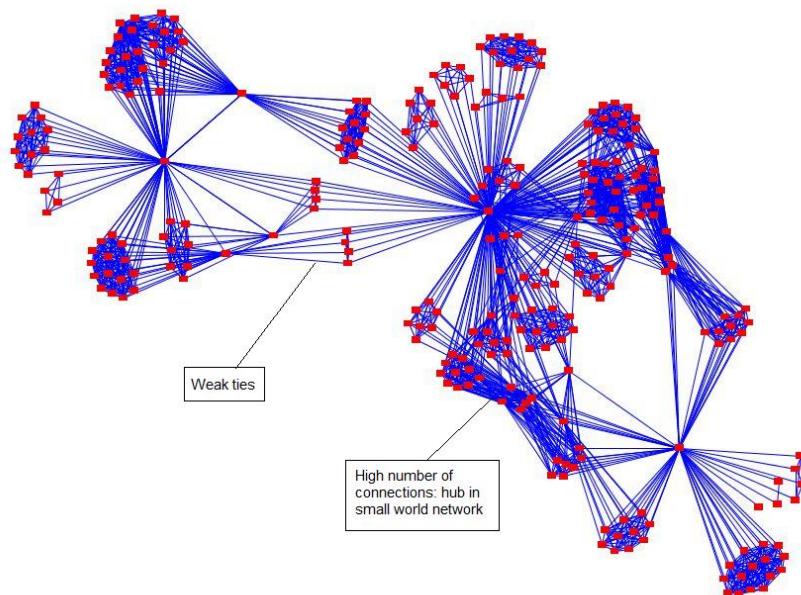


Figure 13.5 – Network formed of small-world and free-scale networks

**Source:** Rozenblat et al. 2013

Some authors, however, (Beauguitte et al. 2011) choose to tone down the contribution of the two concepts to geography, believing that the small-world function's contribution is generally trivial, while that of free-scale has long been known. In contrast, the use of partitioning methods, resulting from research by physicists, has considerably enriched the possibilities for analysing complex networks.

## 13.2 Graph partitioning methods

Whereas a graph makes it possible to represent exchanges between vertices, a partition graph highlights groups of vertices that are connected preferentially. For example, a partition graph showing commercial exchanges makes it possible to indicate where to set up transport platforms to best serve the territory, within each group found by partitioning. Since the 2000s, these methods have been the focus of intensive development efforts, and we can only give here only an introductory view to them, trying to base it on intuitions. They form a branch of graph theory, a fairly long-standing method of analysis (Euler's problem on the bridges of Konigsberg, the problem of map colouring, etc.). Concepts of classical graph theory will be mobilised only when they are essential and we will focus on the concepts specific to large graphs and their partitioning.

### 13.2.1 Concepts in graph theory

**Definition 13.2.1 — Characterising a graph.** A **graph** is a set  $G = \{V, E\}$  (Figure 13.6) in which  $V$  (for *vertex*) refers to a “peak” and  $E$  (for *edge*) the “ridge”.

The **size** of the graph is the number of links.

The **order** of the graph is the number of vertices.

A graph is said to be **empty** when it contains no links.

A graph is said to be **full** when all vertices are connected to all others. There are thus  $\frac{n(n-1)}{2}$  links in a complete order graph  $n$ .

An **oriented** graph is a set of vertices and edges, in which each edge is an ordered pair of vertices. Thus, the relationship between vertices  $x$  and  $y$  is different from that between  $y$  and  $x$ . A **valued** graph, as opposed to a non-valued graph, has multiple links (two vertices are linked multiple times).

In this communication, we will limit ourselves to non-oriented graphs, in which the relations between vertices are symmetrical.

A **simple** graph is one that is not valued and has no loop (no edges that go from a vertex back to the same vertex).

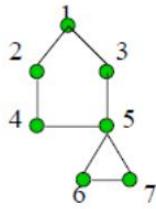
The **degree** of a vertex is the number of vertices with which it is linked. In a simple order graph  $n$ , the degree of a vertex is defined as being between 0 and  $n - 1$ . The sequence of degrees is the series  $d_1, \dots, d_n$ .

The **density** of a graph is the ratio between the number of links observed and the number of links in a complete graph. Thus, it varies between 0 for an empty graph and 1 for a complete graph.

If each point in a graph can be reached from any point, then the graph is **connected** or **connex**.

■ **Example 13.3** The graph shown in Figure 13.6 is a simple graph of size 8 and order 7. ■

While it quickly becomes complex to formalise the theory, some concepts are quite easy to understand. As in spatial statistical methods, an adjacency matrix can be associated with the graph (figure 13.7). A value greater than 0 indicates that there is a link between two points. If the adjacency matrix is symmetrical, then it comes from a non-oriented graph. If the diagonal is equal to 0 then the associated graph is simple (no loop).



(a) A graph with 5 nodes and 8 edges

$$V = \{1, 2, 3, 4, 5, 6, 7\}$$

$$E = \{(1,2), (1,3), (2,4),$$

$$(4,5), (3,5), (4,5),$$

$$(5,6), (6,7)\}$$

(b) In mathematical writing

Figure 13.6 – Geometric and mathematical representations of a graph

0	1	1	0	0	0	0
1	0	0	1	0	0	0
1	0	0	0	1	0	0
0	1	0	0	1	0	0
0	0	1	1	0	1	1
0	0	0	0	1	0	1
0	0	0	0	1	1	0

(a) Adjacency Matrix

2	0	0	0	0	0	0
0	2	0	0	0	0	0
0	0	2	0	0	0	0
0	0	0	2	0	0	0
0	0	0	0	4	0	0
0	0	0	0	0	2	0
0	0	0	0	0	0	2

(b) Degree matrix

2	-1	-1	0	0	0	0
-1	2	0	-1	0	0	0
-1	0	2	0	-1	0	0
0	-1	0	2	-1	0	0
0	0	-1	-1	4	-1	-1
0	0	0	0	-1	2	-1
0	0	0	0	-1	-1	2

(c) Laplace matrix

Figure 13.7 – Adjacency and Laplace matrices associated with the graph in Figure 13.6

A pathway from vertex  $a$  to vertex  $b$  is an ordered sequence of vertices in which each adjacent pair is connected by an edge. One **geodesic** between two points is the minimum length path between these two points. In the example shown in Figure 13.6, a series of vertices  $(1, 3, 5, 7)$  is the geodesic between points 1 and 7, and the series of vertices  $(1, 2, 4, 5, 7)$  is a path, not a geodesic. One point **a** is attainable from a point **b** where there is a path between the two points. If this adjacency matrix is subtracted from the degree matrix (diagonal matrix), the result is a **Laplace** matrix (Figure 13.7 on the right) which plays a fundamental role in the approach referred to as spectral clustering (see *clustering methods*).

All the questions which we will now consider revolve around the possibility of determining, within our graph, sub-graphs referred to as **communities** or **cliques**. This will call forth a discussion of vertices and links which play a particular part, as well as the indicators that make it possible to measure this. Cut-off points and bridges refer respectively to nodes and links, the removal of which reduces the overall connectivity of the graph (Figure 13.8).

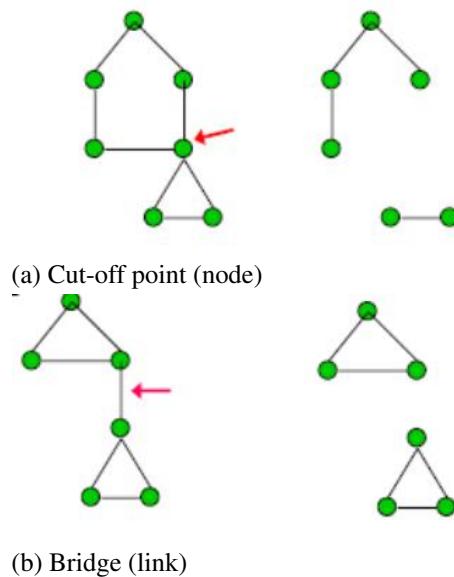


Figure 13.8 – Node or link removal

**Definition 13.2.2 — Some centrality indicators.** Measurements making it possible to consider the most important vertices (and links).

The **connectivity** of a graph is the number of vertices that must be removed to do away with the graph's connected property. Link connectivity is defined in a dual manner, *i.e.*, the number of links to be removed for connectivity to disappear.

Centrality indicators play a very important part in graph analysis and partitioning. Several of them have been defined:

The **term** *degree centrality* simply refers to the degree, *i.e.* the number of links from a vertex. In our example, vertex 5 has the highest degree centrality. This centrality can be standardised by relating it to the number of vertices minus one. This is the simplest of the concepts. It is frequently used in sociology, but does not take into account the structure of the graph.

The **notion of** *closeness centrality* indicates whether the vertex is located close to all the vertices in the graph and whether it can quickly interact with these vertices. It is formally written as

follows:

$$C_c(v) = \frac{1}{\sum_{u \in V \setminus \{v\}} d_G(u, v)} \quad (13.3)$$

with  $d_G(u, v)$  the distance between vertices  $u$  and  $v$ .

The **term betweenness centrality** reflects one of the most important concepts. It measures the utility of the vertex in transmitting information within the network. The vertex plays a central role if many shorter paths between two vertices are to use this vertex. It is written:

$$C_B(v) = \sum_{\substack{i,j \\ i \neq j \neq v}} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \quad (13.4)$$

with  $\sigma_{ij}(v)$  the number of paths between  $i$  and  $j$  passing through  $v$ .

There is also betweenness centrality of links, which reflects the number of geodesics (shortest paths) that pass through a given link. Figure 13.9 shows a link (dark line) with high betweenness centrality. Removing this link leads to the formation of two sub-graphs. This property is used in graph partitioning.

As to **own-vector or spectral centrality**, it is defined by Bonacich from the adjacency matrix. Spectral centrality is a measure of the influence of a node within a network. For a vertex, it is defined as the sum of its connections with the other vertices, weighted by the degree of centrality of those vertices. It can be written as:

$$C(v) = \frac{1}{\lambda} \sum_{u \neq v} A(v, u) C(u) \quad (13.5)$$

which can be written  $\lambda C = AC$ .

To resolve this equation, Bonacich 1987 shows that the spectral centrality vector is actually the dominant (or main) own-vector of the adjacency matrix.

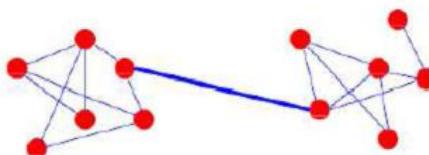


Figure 13.9 – High betweenness centrality (dark line)

It is possible to illustrate these concepts and show how different they are by using one of the most traditional databases, *i.e.* that of Zachary (Zachary 1977) on the social media formed by members of a university karate club (figure 13.10). The *igraph* package in the R software is used to represent the graph and calculate the previous indicators.

---

```
# Degree centrality
d<- degree(kar)
# Closeness centrality
cp<- closeness(kar)
# Betweenness centrality
ci<- betweenness(kar)
```

```
# Own-vector centrality
ce<- graph.eigen(kar)[c("values", "vectors")]


---


kar<-read.graph("karate.gml",format="gml")
plot(kar)


---


```

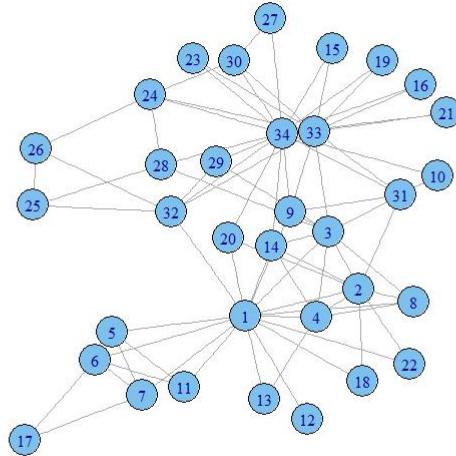


Figure 13.10 – Zachary Network

**Note:** Friendship links between 34 members of a karate club at an American university

The table below shows the ranking of individuals in the network shown in Figure 13.10 according to the different centrality criteria. The ranking is fairly consistent for the first in the ranking. Six individuals share the top five spots on each indicator. Individual 1 is always in the first two positions, particularly for proximity and betweenness. This is owed to Individual 1's large number of links developed (high degree centrality) and role as the necessary intermediary for a small group of individuals (strong betweenness centrality) who are themselves unrelated to the others. Individual 1 is therefore close to all other members of the club, *i.e.* shows high proximity centrality. Own-vector centrality sums up these concepts.

Ranking for each indicator	Degree	Proximity	Betweenness	Own-Vector
First	34	1	1	34
Second	1	3	34	1
Third	32	34	33	3
Fourth	3	32	3	33
Fifth	2	33	32	2

### 13.2.2 Partitioning methods

Back to issues faced with city networks, the first is the determination of communities. In the first chapter, it was shown that city networks often combine "small-world" aspects with strong intra-regional links, and free-scale aspects, with quite highly differentiated sub-groups. We will

base our work largely on the summaries produced by Newman 2006 and Fortunato 2010, as well as on French-language papers by Pons 2007 and Seifi 2012.

### Partition definition and quality

The first problem to address with partition graphs lies in defining the community. No definition is universally accepted. What unifies approaches, without resulting in a precise definition, is that there must be more links within the community than with the rest of the graph. This can only happen if the graphs are low in density – sparse – and if the number of links remains of the same order of magnitude as that of vertices.

The graphs associated with social networks, or some graphs describing biological structures, reach very large sizes, in contrast to those presented so far. Partitioning these graphs into communities requires highly effective algorithms. Their number is growing. They use methods often derived from physics ('greedy' methods, *spinglass*).

As with classification, the issues of community number optimisation, hierarchy and interlocking structures will need to be addressed.

Communities can be approached from a local standpoint, *i.e.* by disregarding as much as possible the graph perceived as a whole. In this spirit, preference will be given to indicators that measure internal cohesion, which could be translated into the language of social networks by the fact that everyone is friends with everyone. In such communities, a large number of **cliques** should appear (complete maximal subgraphs with at least three vertices). Along the same lines, the density of links within the community and that of the links between the community and the rest of the world will also be studied.

They can also be defined by looking at the graph as a whole. One of the main ideas is to compare the structure of a graph showing communities with that of a random graph. These graphs, often referred to as Erdos-Renyi graphs, were the first to be studied. To find more analogies with statistical methods, a *null model* with which to compare our actual graph will be sought. This null model must be a random graph, of course, but one that respects a certain number of constraints in order to be comparable. The most used version is the one proposed by Newman et al. 2004. It consists in a "randomised" version of the original graph, *i.e.* where the links are modified randomly, subject to the constraint that the expected degree of each vertex is that of the original graph. This approach enabled the aforementioned authors to offer one of the most fertile concepts in partitioning theory, *i.e.* that of modularity.

Modularity makes it possible to justify the relevance of the sub-graphs found after partitioning. The strong modularity hypothesis is the comparison with a random graph, implying that a graph with a completely random structure must have a modularity close to 0. This comparison therefore makes it possible to highlight relations that are denser than the average, *i.e.* a community structure, or conversely, if relations are less dense, isolated structures.

**Definition 13.2.3 — Modularity.** This is a measure of the quality of a graph partition. Given partition  $\mathbf{P}$  in  $p$  clusters of graph  $G = \{V, E\}$ , then:  $\mathbf{P} = \{c_1, \dots, c_n, \dots, c_p\}$

Modularity can be introduced quite simply as follows, referring to Newman's idea.

$$Q(P) = \sum_i (e_{c_i} - a_{c_i}^2) \quad (13.6)$$

with  $e_{c_i}$  the percentage of links held by cluster  $c_i$  on the total,  $a_{c_i}$  the probability that a vertex is found in cluster  $c_i$  and therefore  $a_{c_i}^2$  the probability that the two vertices of a link are in the

■ same cluster  $c_i$ .

This general expression is transformed into the first common form of presenting modularity. We show (Fortunato 2010) that modularity can be written as:

$$Q(P) = \frac{1}{2m} \sum_{i,j \in V} \left( A_{ij} - \frac{d_i d_j}{2m} \right) \delta(c_i, c_j) \quad (13.7)$$

with

- $m$  the number of edges in the graph;
- $A$  the graph's adjacency matrix;
- $A_{ij}$  the weight of the links between vertices  $i$  and  $j$ ;
- $d_i$  the sum of the degrees of  $i$  with  $d_i = \sum_j A_{ij}$ ;
- $a_{c_i}^2 = \sum_j \frac{d_i d_j}{4m^2}$ ;
- $\delta(c_i, c_j)$  a Kronecker function worth 1 if the two vertices belong to the same community and 0 otherwise.

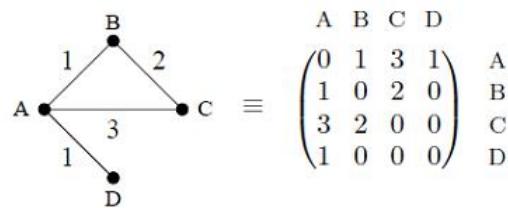
We can show that an alternative way of writing this expression is:

$$Q(P) = \frac{1}{2m} \sum_{k=1}^p \sum_{i,j \in V} \left( A_{ij} - \frac{d_i d_j}{2m} \right) = \sum_{k=1}^p \left[ \frac{l_k}{m} - \left( \frac{d_k}{2m} \right)^2 \right] \quad (13.8)$$

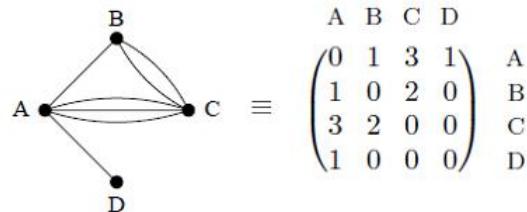
with  $l_k$  indicating the number of links connecting community vertices  $k$  and  $d_k$  the sum of the degrees in community  $k$ .

The term  $A_{ij} - \frac{d_i d_j}{2m}$  reflects the difference in the links between our graph and a random graph where the constraint lies in preservation of vertex degrees.

Modularity definitions were first developed in the context of non-valued graphs. They have since been extended to valued graphs. The value of  $A_{ij}$  is the link between vertices, which in a non-valued graph is worth 1 if the vertices are linked and 0 otherwise, and in a valued graph, the value of the flow if there is a link and 0 otherwise. Newman 2004 offers a very simple way to move from non-valued graphs to valued graphs, introducing what he called multigraphs (Figures 13.11).



(a) Valued graph



(b) Multigraph representation

Figure 13.11 – Multigraphs: switch to valued graphs

This representation makes it possible to extend the results presented above to weighted graphs.  $A_{ij}$  are the weights connected with the links or equivalent to the number of links in the multigraph.  $M$  is the number of links in the multigraph, or the sum of the weightings.

Modularity is one of the most powerful concepts in the graph partition theory. Although it has been criticized, it is still the most used. It is used as a foundation for certain methods, and as a measure of the quality of the partitions produced using other methods. It will be used several times in the examples given further.

The work of Guimaraes, Reichart and Bornholdt, highlighted by Fortunato (Fortunato 2010) look at the problem of "resolution". If the number of links in the graph becomes very large and the expected number of links (see modularity formula) is less than 1, a single link between the two groups is enough to merge them.

### General overview of partitioning methods

Once the general scheme of a partition has been determined, it remains to be put into practice. Concretely, this means finding ways of proceeding, *i.e.* algorithms, which, first of all, make it possible to solve the problem, then actually solve it within an acceptable time. City network graphs are already large but are very small when compared to social networks or even to those used in protein or genome studies. The complexity of algorithms (NP-hard or NP-complete problems) is presented in Fortunato 2010. Researchers often attempt to measure the complexity of algorithms by noting them as  $O(n^2m^2)$  with  $n$  the number of links, and  $m$  the number of edges. The methods take up well-known data analysis questions — how many classes are there? should they be determined beforehand? should bottom-up or top-down methods be applied? how should the stopping criteria be determined? In this presentation, we will cover only a few families of methods tested as part of the research carried out by INSEE's "Territorial Analyses" Division on city networks (see section 13.3), focusing on those implemented in the R software. The methods are rapidly developing at present and are subject to controversy among specialists. Many of those presented here come from the work of Mark Newman, who originated, among other things, the concept of modularity presented in the previous paragraph. The questions' algorithmic complexity has led to a great deal of initial research on graph bi-partition (Kernighan et al. 1970). Other methods were also inspired by previous research on data analysis (classification dendograms, *k-means* methods). Such methods are based on the properties of the graphs, or on the treatment of the adjacency matrix.

### Classic methods

We will present only a few of the classic methods:

#### Methods based on graph bi-section

These methods (Figure 13.12) are quite simple to present. The idea is to search for the line that splits the graph by cutting out the lowest possible number of links (*cut size*).

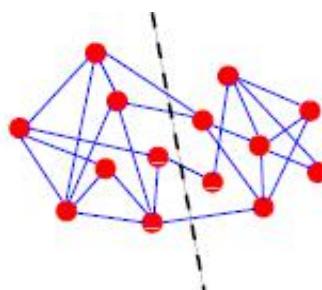


Figure 13.12 – Graph bipartition

However, this method – in its simplest version – runs the risk of showing only trivial solutions (an isolated vertex). More elaborate bisection methods are based on spectral methods (properties of the Laplace matrix spectrum) which will be presented below.

### Hierarchical methods

These methods (Figure 13.13) are based on measures of similarity between vertices. Once we have calculated similarity for each pair of vertices (similarity matrix), we can for instance build a dendrogram using fairly classic methods.

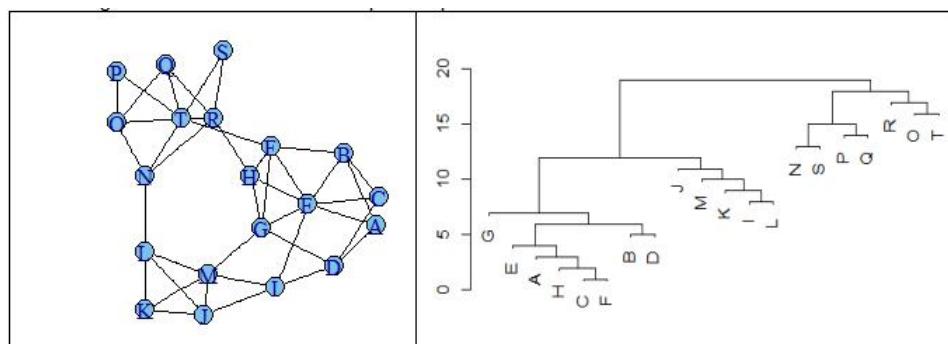


Figure 13.13 – Hierarchical partitioning methods

### Clustering methods

These methods are well known in data analysis. In these methods, the number of classes is predetermined. A distance between couples of points is defined, longer if the vertices are dissimilar. The aim will be to minimize a cost function based on points and centroids. As a minimum *k-means clustering*, for example, the cost function is the longest distance between two points of the class. Our aim will be to find the partition that minimises the largest of the  $k$  classes (search for compact classes). The MacQueen method is based on minimizing total intra-class distances.

#### **The division method**

This method is one of the most intuitive to present. It is based on the concept of betweenness centrality presented in section 13.2.1, with a diagram which presents this idea quite well in a simple case. When many geodesics from one point of the graph to another pass through a vertex or a link, removing them is more likely to bring out communities. In the example shown above (Figure 13.14), the link between vertices  $T$  and  $F$  has the highest betweenness centrality. If this link is deleted, then the  $RH$  link has the highest centrality, followed by the  $NL$  link.

After removing these three links, the graph is no longer connected and a community appears. The process can continue. An R command produces the final outcome.

---

```
karate <- read.graph("karate.gml",format="gml")
plot(karate,vertex.size=2)
betkar<- edge.betweenness.community(karate)
plot(betkar,karate)
```

---

The result on this very simple graph is quite trivial and reflects what it produces on a graph that is still readable, but more complex, *e.g.* that of the karate club. The best known division method is that of Newman et al. 2004. It also confirms the appeal of studying graphs for physicists. The algorithm illustrated above is as follows:

1. calculate betweenness centrality for all links;
2. remove link with the highest centrality;

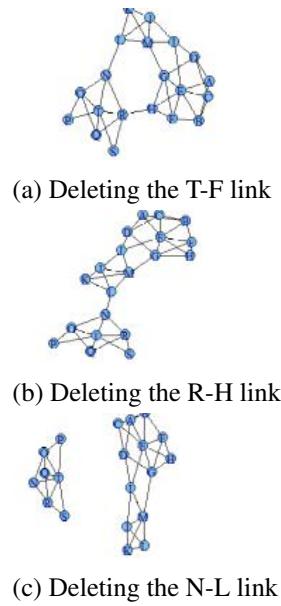


Figure 13.14 – Graph partition in Figure 13.13 using the division method

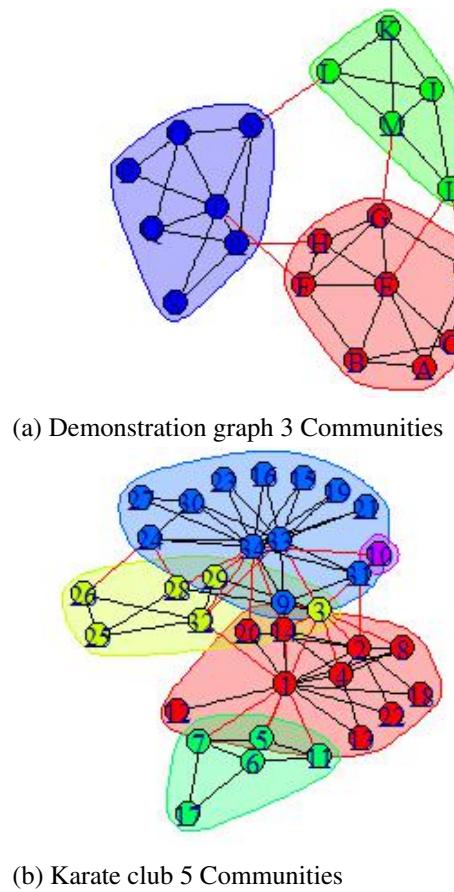


Figure 13.15 – Result of the division method on a graph and on the karate club

3. re-calculate centrality;
4. iterate the cycle in step 2.

This iterative process can continue until all the vertices are isolated, thus producing a hierarchy of interlocked partitions. The partition can be chosen using the modularity criterion. This algorithm requires, at each stage, that the betweenness centralities be calculated, and its complexity is in  $O(m^2n)$ , making it unusable on very large graphs.

Other division algorithms have been proposed. Fortunato 2010 proposed an algorithm that uses the information centrality link defined as the relative decrease in network efficiency when this link is removed from the graph. This algorithm is more efficient, but more complex than that of Girvan-Newman. The latter therefore remains widely used, in particular as a comparison of the communities detected.

### **Agglomerative methods based on modularity**

This family of methods is very rich and very important. In contrast to the division method, it starts from all the vertices, gradually aggregating them.

#### The “optimal” method

It is based on exploring all possible communities and maximising modularity. The work produced by Fortunato 2010 offers a value approaching the number of these partitions, which explodes with the size of the graph and makes it unusable, even for medium-sized graphs. Calculating communities from this perspective uses a physics-based method referred to as “simulated annealing” (successive heating and cooling in a state of equilibrium), which is often used in optimisation issues. It is implemented in R in the *igraph* package by the `optimal.community` command.

#### The Clauset and Newman Method - Aggregative Method

The algorithm is said to be ‘greedy’ in that it makes it possible to create a partition based on a modularity criterion. It was first proposed by Newman in 2003, then by Clauset, Newman and Moore in a second version. It uses modularity in the following form:  $Q = \sum_i (e_i - a_i^2)$ . A magnitude stated as  $\Delta Q_{ij}$  is defined, reflecting the change in modularity when linking community  $i$  and community  $j$ . Details of the algorithm, along with instructions for information storage, can be found in Clauset et al. 2004. The general diagram is as follows:

1. The process starts with  $n$  communities (each vertex being a community);
2. for each pair,  $\Delta Q_{ij}$  is calculated;
3. the pairs that most increase modularity are merged;
4. phases 2 and 3 are repeated until obtaining a single community;
5. the dendrogram is cut to the value that reflects the highest modularity.

In this very simple example (figure 13.16), modularity  $Q$  can be seen as increasing up to stage 10, where the three fairly visible communities are identified. In stage 11, two of the communities merge and modularity decreases, becoming null when the three communities are combined. The result is therefore a partitioning into three communities with a modularity of 0.485. This algorithm is implemented in R in the *igraph* package by the `fastgreedy.community` function. The characteristic of this algorithm is its high execution speed, which enables it to be used on large graphs. The algorithm is of complexity  $O(mn)$ .

#### Spectral methods

Newman has proposed a spectral version of partitioning based on modularity. In this version, the matrix introduced shows the modularity expression  $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$ .

In the initial case of a bipartition, which was later generalised, Newman introduced a vector  $s$  worth +1 if the vertex belonged to the first group, (1) if it was part of the second. It shows

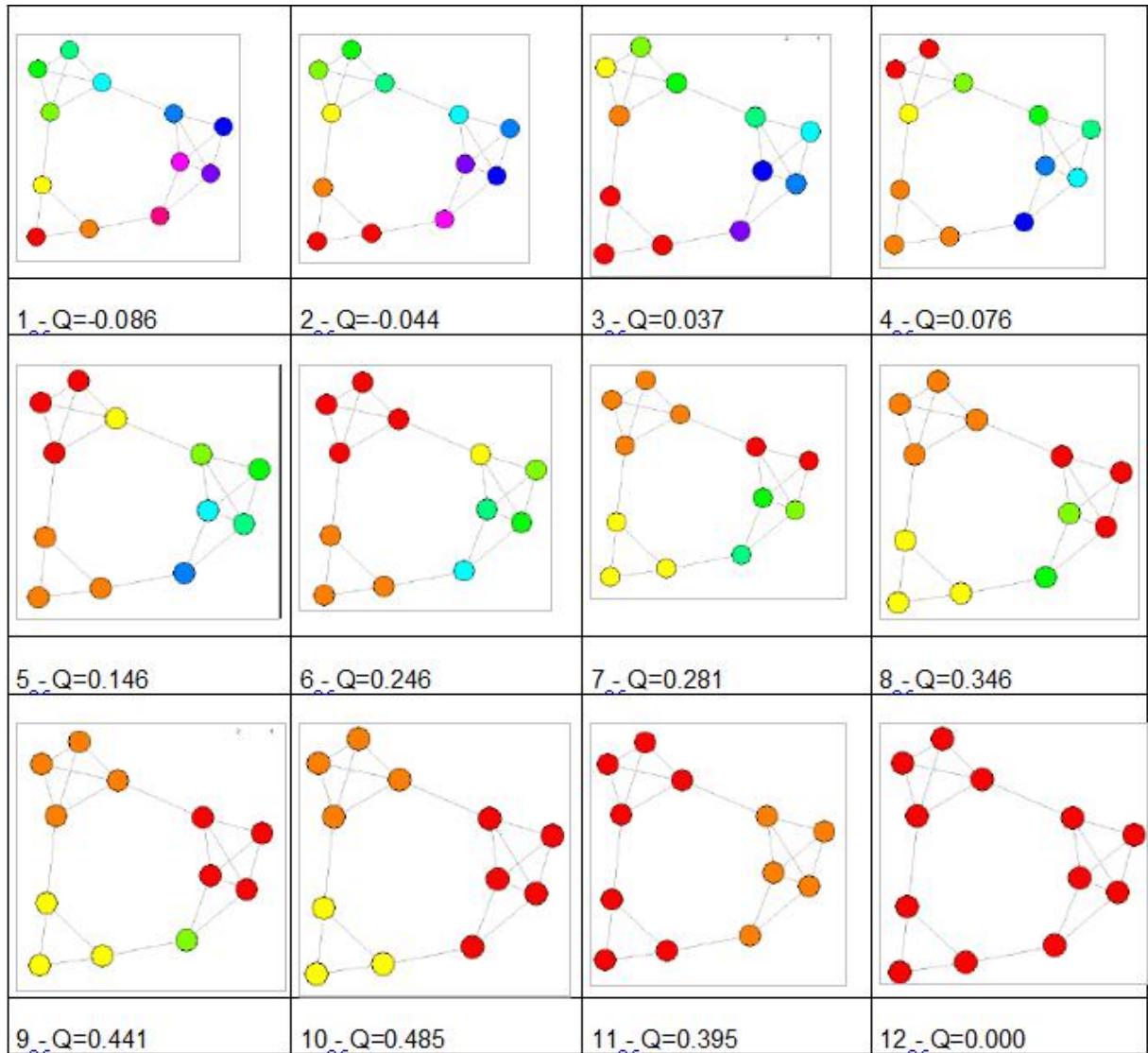


Figure 13.16 – The 12 stages of partitioning a 12-vertex graph using the aggregative method

that the maximisation of modularity according to vector  $s$  is a problem that can be formalised using  $B_s = \lambda D_s$  in which  $\lambda$  is a Lagrange multiplier, and  $D$  a diagonal matrix containing vertex degrees. When this matrix problem is solved, given the structure of the matrix on which we work, the result is a trivial solution with an own-value equal to 0 and a vector composed of 1, or bringing all vertices together in a single community. To carry out the partitioning, the clean vector associated with the highest own-value is used (Newman 2006). The *igraph* package contains the `leading.eigenvector.community` function which implements this method.

### Louvain Algorithm

In 2008, three researchers from the University of Louvain proposed another "greedy" method, which was faster than most other approaches. It is distinctive in that it is based on a local approach to modularity. In the first phase, a different community is attributed to each vertex. Next, the neighbours of each vertex  $i$  are considered, and the modularity gain is calculated by removing vertex  $i$  and placing it in community  $j$ . A positive and maximum gain is needed to move  $i$ . This is done sequentially until no improvement is possible. The second phase of the algorithm consists in building a new network whose vertices are the communities identified in the first phase, the weights of the links between the communities being determined by the sum of the weights of the links at the vertices of the initial graph. Once this second phase has been completed, the algorithm is re-applied to this new weighted network. A combination of the two phases is a "pass", and these passes are iterated until maximum modularity is reached. The *igraph* package contains the `multilevel.community` function which implements this method. It is often presented, particularly in Newman's recent articles, as the most effective in terms of time and partitioning quality (Newman 2016).

## Other methods

### Random walks

The `walktrap.community` algorithm ultimately aims, like all others, to produce distances between the vertices of the graph. The idea is to reach this distance based on the idea of random walking. Time becomes discrete. At all times, a walker moves randomly from one vertex to another vertex chosen amongst its neighbours. The series of vertices visited is then a random walk. The probability of going from vertex  $i$  to vertex  $j$  is:

$$P_{ij} = \frac{A_{ji}}{k_i}. \quad (13.9)$$

The transition matrix of the corresponding Markov chain is thus found, and  $P_{ij}(t)$  – the probability of passing from vertex  $i$  to vertex  $j$  in a time  $t$  – can be calculated. When a random walk in a graph is long enough, the probability of being on a given vertex is directly (and solely) proportional to the degree of that vertex. The probability of going from  $i$  to  $j$  and that of going from  $j$  to  $i$  by a random walk of fixed length have a proportionality ratio that depends only on the degrees of the start and end vertices:

$$k_i P_{ij}(t) = k_j P_{ji}(t). \quad (13.10)$$

The method used for comparing two vertices  $i$  and  $j$  must be based on the following findings:

- if two vertices  $i$  and  $j$  are in the same community, then probability  $P_{ij}(t)$  is most likely high. However, if  $P_{ij}(t)$  is high, it is not always guaranteed that  $i$  and  $j$  are in the same community;
- probability  $P_{ij}(t)$  is influenced by the degree of  $k_j$ , the arrival vertex. Random walks are more likely to pass through high-degree vertices (in the case of limitless random walking, this probability is proportional to the degree);

- vertices of the same community tend to see vertices similarly distant, so if  $i$  and  $j$  are in the same community and  $k$  in another community, there is a good chance that  $P_{ik}(t) = P_{jk}(t)$ . This defines a distance, which must be lower when the two vertices belong to the same community:

$$\sqrt{\sum_{k=1}^n \frac{(P_{ik}(t) - P_{jk}(t))^2}{k_k}}. \quad (13.11)$$

In this method, the choice of  $t$  is very important. If  $t$  is too small, the communities are tiny. If too large, the probabilities tend towards the same value. Once the distance matrix has been determined, the algorithm is quite classic — we start from  $n$  communities and then we aggregate. A tree is obtained and modularity is used to find the appropriate partition. Details can be found in Pons 2007.

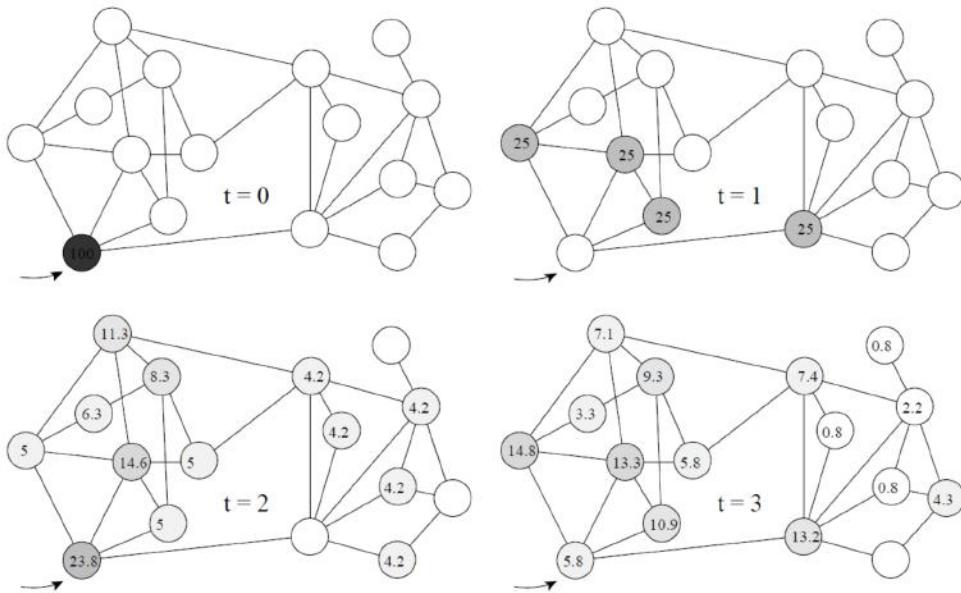


Figure 13.17 – Illustration of random walking on a graph

**Source:** according to Pons 2007

In the example shown in Figure 13.17, we have – up to  $t = 3$  – graphically represented the probability matrix that will be used for the partitioning (by spectral analysis). The *igraph* package offers the `walktrap.community` function that implements this method.

### Spin glasses

This method is a move away from the usual methods. It is inspired by spin glasses, which are impurities alloys, with a spin being associated with each impurity. The coupling between the different spins can be of varying degrees of intensity. This method is used in theoretical physics. Spin pairs are associated in a graph. A **Hamilton graph** (graph with at least one cycle passing through all vertices at most) is defined and probability distribution of couplings is set. Reichardt et al. 2006 have used this approach. Each vertex is characterised by a spin with  $q$  possible values, while the communities are made up of the vertex values with equal spin values. The energy of the system is defined by a hamiltonian using the graph's adjacency matrix. This expression is minimised by simulated annealing, as in the case of the “optimal” method presented previously. The *igraph* package offers the `spinglass.community` function that implements this method.

## References - Chapter 13

- Barabási, Albert-László and Réka Albert (1999). « Emergence of scaling in random networks ». *Science* 286.5439, pp. 509–512.
- Battiston, Federico, Vincenzo Nicosia, and Vito Latora (2014). « Structural measures for multiplex networks ». *Physical Review E* 89.3, p. 032804.
- Beauguitte, Laurent and César Ducruet (2011). « Scale-free and small-world networks in geographical research: A critical examination ». *17th European Colloquium on Theoretical and Quantitative Geography*, pp. 663–671.
- Bonacich, Phillip (1987). « Power and centrality: A family of measures ». *American journal of sociology* 92.5, pp. 1170–1182.
- Christaller, Walter (2005). « Les lieux centraux en Allemagne du Sud Une recherche économico-géographique sur la régularité de la diffusion et du développement de l'habitat urbain ». *Cybergeo: European Journal of Geography*.
- Clauset, Aaron, Mark EJ Newman, and Cristopher Moore (2004). « Finding community structure in very large networks ». *Physical review E* 70.6, p. 066111.
- Fortunato, Santo (2010). « Community detection in graphs ». *Physics reports* 486.3, pp. 75–174.
- Karinthy, Frigyes (1929). « Chain-links ». *Everything is the Other Way*, p. 25.
- Kernighan, Brian W and Shen Lin (1970). « An efficient heuristic procedure for partitioning graphs ». *The Bell system technical journal* 49.2, pp. 291–307.
- Newman, Mark EJ (2004). « Analysis of weighted networks ». *Physical review E* 70.5, p. 056131.
- (2006). « Modularity and community structure in networks ». *Proceedings of the national academy of sciences* 103.23, pp. 8577–8582.
- Newman, Mark EJ and Michelle Girvan (2004). « Finding and evaluating community structure in networks ». *Physical review E* 69.2, p. 026113.
- Newman, Mark, Albert-Laszlo Barabasi, and Duncan J Watts (2011). *The structure and dynamics of networks*. Princeton University Press.
- Newman, MEJ (2016). « Community detection in networks: Modularity optimization and maximum likelihood are equivalent ». *arXiv preprint arXiv:1606.02319*.
- Pons, Pascal (2007). « Détection de communautés dans les grands graphes de terrain ». PhD thesis. Paris 7.
- Reichardt, Jörg and Stefan Bornholdt (2006). « Statistical mechanics of community detection ». *Physical Review E* 74.1, p. 016110.
- Rozenblat, Céline and Guy Melançon (2013). *Methods for multilevel analysis and visualisation of geographical networks*. Springer.
- Seifi, Massoud (2012). « Cœurs stables de communautés dans les graphes de terrain ». PhD thesis.
- Watts, Duncan J and Steven H Strogatz (1998). « Collective dynamics of 'small-world' networks ». *nature* 393.6684, p. 440.
- Wilson, Alan Geoffrey (1974). *Urban and regional models in geography and planning*. John Wiley & Sons Inc.
- Zachary, Wayne W (1977). « An information flow model for conflict and fission in small groups ». *Journal of anthropological research* 33.4, pp. 452–473.



# 14. Confidentiality of spatial data

MAËL-LUC BURON, MAËLLE FONTAINE  
INSEE

---

<b>14.1 How to evaluate spatial disclosure risk?</b>	<b>351</b>
14.1.1 General definition of disclosure risk . . . . .	351
14.1.2 Specificities of spatial data regarding disclosure risk . . . . .	351
14.1.3 Recommendations to measure disclosure risk of spatial data . . . . .	353
<b>14.2 How to deal with disclosure risk?</b>	<b>354</b>
14.2.1 Pre-tabular or post-tabular SDC methods? . . . . .	355
14.2.2 Overview of SDC methods taking geography into account . . . . .	355
14.2.3 How to assess effectiveness of disclosure control? . . . . .	359
<b>14.3 Application for a grid of 1 km<sup>2</sup> squares</b>	<b>360</b>
14.3.1 Targeting Record Swapping: details of the method . . . . .	361
14.3.2 Choice of data and parameters . . . . .	362
14.3.3 Results . . . . .	364
<b>14.4 Differencing issues</b>	<b>367</b>
14.4.1 Definition . . . . .	367
14.4.2 Illustration . . . . .	367
14.4.3 Identifying Risky Areas . . . . .	368
14.4.4 Protection methods . . . . .	369

---

## Abstract

Recent profusion of geocoded sources, often released as grid data, allows many possibilities of analysis for economists, demographists, or sociologists. But it also leads to high disclosure risk, because the number of variables necessary to uniquely identify someone considerably decreases when the intruder knows the geographical location at a detailed level. This issue is even more serious in areas with low population density. Traditionally, Statistical Disclosure Control (SDC) methods do not take spatial features of data into account. This chapter aims at giving suggestions to measure and to deal with disclosure risk, while preserving spatial correlations. Pre-tabular methods seem to be more appropriate to the purpose because they can target the riskiest records, with a measure of risk depending on the local context. But applying only pre-tabular methods cannot lead to a sufficient level of protection, and post-tabular methods can be performed in a second step to guarantee respect of national regulations. In this chapter, we give an overview of the existing literature and we make a focus on a specific SDC method called Targeted Record Swapping (TRS), highlighted by the Eurostat Grant "*Harmonised Protection of Census Data in the ESS*". This method detects the records most exposed to disclosure and swaps them with other similar records belonging to the same geographical neighbourhood. Therefore, individuals with rare attributes continue to be present in the data, but not at their actual location. This ensures that an intruder cannot re-identify them with certainty. We test TRS on French fiscal data for a small region and for several

parameters, and we obtain very little distortion of spatial correlations for variables taken into account in the method or strongly correlated to them.



Prior reading of Chapters 2: "Codifying the neighbourhood structure" and 3 "Spatial autocorrelation indices" is recommended.

## Introduction

The analytical richness of spatial data and how they can explain underlying phenomena has been widely commented in previous chapters. Analytical tools to take advantage of this wealth of information have also been presented.

In the near future, more and more data will be geolocated, leading to a profusion of information available at a very detailed geographical level and giving to economists and analysts many topics to explore. But this profusion also leads to crucial stakes concerning confidentiality of spatial data. The number of characteristics necessary to uniquely identify an observation decreases with the size of the mesh within which information is released, and even more in the context of proliferation of open access geographical visualisation tools. When population density is low in a given area, disclosure risk increases, because the probability to find someone else similar in the neighbourhood is low.

The situation is a conflict between two great principles of public statistics releasing (VanWey et al. 2005). On the one hand, National Statistical Institutes (NSIs) have the vocation of offering as much data as possible with a high level of utility, and on the other hand, they have to manage with strong constraints to guarantee and enforce the confidentiality of information providers. Ensuring confidentiality of spatial data is indeed a particularly difficult task because European and national regulations prohibit NSIs from disseminating any data that could allow an intruder to identify, directly or indirectly, the investigated household or company, and scrupulously that would mean, in most cases, not to release anything. Finally, any data releasing means a non-zero disclosure risk, and the stake is to reduce it to a low and acceptable level. In other words, data protection strategy can be seen as a trade-off between minimising disclosure risk and maximising data utility.

This chapter is not written from the point of view of the user of confidential data; it is written from the point of view of the expert in statistical disclosure control, whose task is to disseminate data satisfying statistical confidentiality regulations, conditionally on some output strategies. Typically, he has a file of individual data (micro-data), and he has to disseminate tabular data for small regional breakdowns or grid data from it, but it is forbidden to disseminate a statistic if it concerns less than a certain threshold of observations. In what follows, a record stands for a household, an individual, or an organisation. We assume the micro-data to be exhaustive: the methods presented are not valid for survey data.

Section 14.1 introduces the disclosure risk: how it can be defined in case of spatial data, and what recommendations can be made to detect high risk observations. Eurostat published in 2007 a Handbook on statistical control (second version in 2010) about standard methods used to deal with confidentiality issues, but spatial data requires adaptations regarding confidentiality treatment. Section 14.2 gives an overview of different methods to deal with disclosure risk for spatial data, including recommendations about risk-utility analysis. Section 14.3 presents the results of some pre-tabular methods tested on a French region, in the context of grid data releasing. For these tests, differencing issues with administrative zoning are not treated, but Section 14.4 will specifically focus on the subject.

## 14.1 How to evaluate spatial disclosure risk?

### 14.1.1 General definition of disclosure risk

Maintaining privacy is essential to retaining the trust of providers, with in the background, the fear of falling response rates. A respondent to a survey must be confident that his personal information will be safeguarded. European regulations have been written to coerce confidentiality<sup>1</sup>: according to Article 20, Chapter V of Commission Regulation No 223/2009 on European Statistics: "*Within their respective spheres of competence, the NSIs and other national authorities and the Commission (Eurostat) shall take all necessary regulatory, administrative, technical and organisational measures to ensure the physical and logical protection of confidential data (statistical disclosure control).*". Countries also have their own regulations. Confidentiality constraints usually take the form of adequate thresholds: no information can be disclosed if it concerns less than a given number of observations. The choice of thresholds depends on various parameters: sparsity of the area, risk aversion, sensitivity of variables, future data users. Sometimes, guidelines are available to check if the data file complies with the confidentiality rules (ONS 2006, Insee 2010).

Disclosure occurs when an intruder (also called "data snooper" in some articles) uses released data to learn some information he did not already know. The intruder is not a hacker. He just has released data at his disposal, and he does not attempt to break any security system. A distinction can be made between different disclosure scenarios (Duncan et al. 1986, Lambert 1993, Clifton et al. 2012<sup>2</sup>, Bergeat 2016):

- **identity disclosure** occurs when a direct identifier of a statistical individual (company, household or person) can be found thanks to the released data (for example, it can be easy to identify the company of a sector with the most important turnover);
- **attribute disclosure** occurs when the intruder can reveal some association between a respondent and some of his sensitive variables ("quasi-identifiers"). Identity disclosure always implies attribute disclosure but the opposite is not true. For example, if the intruder knows someone living in an area, and if the released data show that all the inhabitants of this area share a common characteristic, then he can deduce that the individual has the characteristic, even if he cannot deduce his other attributes;
- **inferential disclosure** occurs when an intruder can infer some attribute with high confidence. Generally, this type of disclosure is not taken into account to protect a dataset.

To comply with strict regulations, one approach is to consider different kinds of users. General users will only have access to less information (less variables or larger categories), whereas specific public like researchers will have restricted access to more data in secure centres, if they previously justify their request by some procedure.

A complementary approach is to introduce perturbation in the data, in order to reach an acceptable level of disclosure risk. Applying a statistical disclosure control (SDC) method then consists in reducing data utility in exchange of more protection. Traditionally, SDC methods do not take spatial features into account, and spatial correlations can be very distorted before and after perturbation. The next subsection gives arguments in favour of geographically intelligent SDC strategies.

### 14.1.2 Specificities of spatial data regarding disclosure risk

Disclosure control experts dealing with spatial data face a paradox. On the one hand, such data need more protection because they permit more identifications, but on the other hand they offer many possibilities of analysis that users do not want to distort too much.

1. With equivalents outside Europe like Australian Privacy Act in 1988  
2. Clifton et al. 2012 makes a classification of different disclosure risks and defines first-tier, second-tier et third-tier.

### Theoretical considerations

In the handbook of SDC guidelines from Eurostat (Hundepool et al. 2010), three different levels of quasi-identifiers are suggested. Only geographical location is considered in the category of extremely identifying variables. Disclosure risk is indeed higher when considering spatial data, for several reasons.

Firstly, the risk of identity disclosure increases in presence of spatial data because it is easier to mobilise personal knowledge. Indeed, among the characteristics possibly shared with someone (age, gender, etc.), belonging to a neighbourhood leads to a higher probability of personally knowing the person. Moreover, it has recently become possible to identify addresses with the development of web scraping or open access tools like *Google Earth*, that make possible re-engineering (Curtis et al. 2006) or direct identification (Elliot et al. 2014). As a result, population density is a fundamental predictor of disclosure risk: the lower the density, the higher the disclosure risk.

Secondly, the risk of attribute disclosure increases in case of spatial data because of Tobler's first law of geography, which states that "*everything interacts with everything, but two close objects are more likely to do so than two distant objects*". As a result, the degree of dissimilarity of an individual to his neighbours seems to be another good predictor of disclosure risk.

And finally, disclosure risk increases with the differencing issue. When data is disseminated in different non-nested geographies (typically administrative borders and grid of squares), in some cases, someone's attribute can be deduced by subtracting the counting of an area from the counting of another enclosing area. Therefore, anyone proficient with geographic information systems (and subtractions) becomes a potential intruder. This particular question of geographic differencing is the topic of Section 14.4.

### Technical considerations

Technically, the dissemination classification (zoning, administrative boundaries, or regular lattice like a grid of squares), is a categorical variable like any other (additional dimension of tabulated data). It is therefore possible, with classical software, to deal with disclosure risk without any geographical consideration, simply considering the mesh as a variable with many modalities. Nevertheless, a geographically intelligent management of disclosure issues will preserve underlying spatial phenomena, but no specific software has been developed yet.

In practical terms, dealing with spatial data adds a layer of complexity in the disclosure control process because it requires much computing power. On micro-data, some SDC methods involve specifying the neighbourhood structure with a weight matrix (so-called "W matrix"), whose size can easily become unmanageable for classic computers. On tabular data as well, detecting the risky observations by differencing sometimes requires to combine many dimensions (NP-hard issue).

### A growing preoccupation

Last but not least, the specificity of spatial data is that they have become more and more numerous and popular, especially in the form of grid data. Increasing geolocation of data by NSIs make it possible to disseminate increasing amount of grid data (at national or global level<sup>3</sup>).

Grid data has many advantages. It brings a satisfactory answer to the need of having a better representation of socio-economic realities and getting rid of administrative zoning, that do not reflect socio-economic or natural realities (Clarke 1995, Deichmann et al. 2001). It gives a better description of sparse areas, like in Finland and Sweden (Tammilehto-Luode 2011). Since the

3. In the 1990s, the project "*Gridded Population of the World*" began to apply these principles to global geography. It has been followed by a continuous improvement of the resolution (Deichmann et al. 2001). In the beginning of 2010, the Geostat project was launched in cooperation between Eurostat and the European Forum for Geography and Statistics (EFGS). The first part of the Geostat project dealt more specifically with grid data (Backer et al. 2011), and the second part aimed at fostering a better integration of statistics and geospatial information in order for the statistical community to provide more qualified descriptions and analyses of society and environment (Haldorson et al. 2017).

squares have always has the same size, grid data ensures comparability over territories and time. If needed, squares can be reassembled to form customisable study areas. Grid data also constitute a good source for auxiliary data or for local sample. Finally, it is easy to integrate data of different nature to grid data, with possible use cases in many disciplines: meteorology, environment, health, telecommunications, marketing, etc.

The next section presents how, in this context of profusion, geographical concerns can be introduced into the choice of SDC methods to keep maximum data utility for geographical analysis.

#### 14.1.3 Recommendations to measure disclosure risk of spatial data

Quantitatively evaluating disclosure risk is a crucial step for SDC experts. With non spatial data, disclosure risk metrics have been developed and discussed (Willenborg et al. 2012, Duncan et al. 2001, Doyle et al. 2001). They are often based on a decision-theoretic characterization of the intruder (Lambert 1993, Duncan et al. 2001). To describe the final micro dataset,  $k$ -anonymity and  $l$ -diversity are common concepts. A dataset satisfies  $k$ -anonymity if for each combination of values of quasi-identifiers there are at least  $k$  observations, whereas the dataset satisfies  $l$ -diversity when for each combination of quasi-identifiers there are at least  $l$  "well represented" values for sensitive attributes. The  $l$ -diversity model extends the concept of  $k$ -anonymity with intra-group diversity for sensitive attributes in order to prevent group disclosure through homogeneity.

With spatial data, individual measures of risk can be calculated to take into account the fact that a record is risky conditionally to a geographical level or to the personal knowledge of the intruder. But the task is not easy because there is no consensual binary measure of risk.

Whether the data is spatial or non-spatial, one approach is to build the tabular data just as if it were disseminated without any constraint, and to flag risky cells as cells that do not satisfy the constraints (minimum cell sizes, dominance rule (also called (n,k) rule), p% rule<sup>4</sup>). Then risky records are all the records inside risky cells. For grid data or small mesh data, risky areas can be flagged with these same rules, considering the mesh or the square like a dimension like another of the tabular data.

Another approach, appropriate for pre-tabular methods, is to work directly on the micro-data. Each observation is associated to a probability of being reidentified by an intruder. The underlying idea is that an observation is risky if it is not surrounded by similar observations. Conditionally to a list of quasi-identifiers, a score evaluates, for each record, the likelihood to find someone else sharing the same characteristics in the neighbourhood. An individual alone in an empty area will always be considered as risky, but an elderly man located in an area with mainly young people will be risky as well.

Ideally, such a score requires choosing a definition of distance or neighbourhood between two records (euclidean distance, number of households in a circle, queen or rook neighbourhood<sup>5</sup>), and to build a  $n \times n$  matrix on exhaustive data<sup>6</sup>. For populous areas, this computation quickly encounters computing power issues. To solve this, an alternative is to base the risk measure on:

- frequency counts of sensitive variables (see also the "special uniques" algorithm developed in Elliot et al. 2005);
- a simpler definition of the neighbourhood: belonging to a same area at a superior hierarchical level. That supposes to have a nested system of geographical levels<sup>7</sup>. In this case, spatial location of the records is not directly used.

4. All these rules are well-known in disclosure control literature and will not be developed here.

5. See Chapter 2.

6. Where  $n$  is the number of records in the micro-data

7. This hierarchical system can be the final releasing support, or can be drawn specifically.

Two examples to target the SDC method to the riskiest records are presented below.

**Box 14.1.1** In Shlomo et al. 2010, a score is calculated for each record as follows.  $M$  key variables (quasi-identifiers, all categorical) are selected, each having  $k_m$  categories ( $m = 1, \dots, M$ ). We are in a hierarchical system of geographical levels (for example nested NUTS partition, or grid of squares of different sizes). For each geographic level  $l$  with  $G$  modalities ( $g = 1, \dots, G$ , for example  $G$  squares), the univariate frequency count is denoted  $N_k^{g,m}$  ( $k = 1, \dots, k_m$ ). The table of  $N_k^{g,m}$  below has  $G * \sum_{m=1}^M k_m$  cells.

$g$	Mod. A1	Mod. A2	Mod. A3	Mod. B1	Mod. B2
1	5	4	1	7	3
2	4	3	3	9	1
...					
G	5	0	5	6	4

For each level of geography  $l$  (for example for the square level), Shlomo et al. 2010 calculate for every record  $i$  (having modalities  $(k_1^i, \dots, k_M^i)$  and belonging to the mesh  $g^i$ ) a score equal to the average of the reciprocal counts:

$$R_i^l = \frac{\sum_{m=1}^M 1/N_k^{g_i^l, m}}{M} \quad (14.1)$$

In the example above, an individual  $i$  in region  $g=1$  taking modalities (A1, B1) will have a risk equal to  $(1/5 + 1/7)/2 \simeq 0.17$ . Then thresholds  $T^l$  are set for each level of geography  $l$ , and scores above thresholds flag risky records. Thresholds generally correspond to quantiles; they are set by the expert who decides which proportion of the population must be considered as risky, with the problem of being data-specific<sup>a</sup>.

Hungary Census grid-based Statistics use the same approach to target risky individuals but introduce multivariate distributions: in Nagy 2015, flag values are calculated for every possible combination of 3 chosen attributes (including the grid square), and the  $n$  riskiest records will be the  $n$  first, after sorting the micro-data by decreasing sum of flag values. The number of cells in the frequency counts table is then  $G * \prod_{m=1}^M k_m$  cells (sparse table). If  $M$  is high (or if most of  $k_m$  are high) then computation power issues can be encountered. A solution can be to create *ad hoc* quasi-identifiers crossing relevant variables, or to add *a posteriori* to the at-risk sample records with very rare combination of modalities (like widows under 20 years old).

a. A threshold can be relevant for some size of mesh but irrelevant for another. For example, 10% of risky squares does not mean the same thing if the mesh is 10 meters or 10 kilometres.

In all cases, high risk households are generally defined as any household having at least one high risk record.

## 14.2 How to deal with disclosure risk?

Now risky records have been identified, perturbation must be added to the data, in order to make the global level of risk acceptable. Section 14.2.1 gives generalities about disclosure control techniques, and Section 14.2.2 gives an overview of SDC methods taking specificities of spatial data into account. To finish, Section 14.2.3 suggests tools and metrics to assess the effectiveness of the chosen method.

### 14.2.1 Pre-tabular or post-tabular SDC methods?

Traditionally, in disclosure control literature, a distinction is made between post-tabular methods, applied on tables (hypercubes) and pre-tabular methods applied on micro-data. Concerning census data, in practical terms, most countries adopt post-tabular methods, for example aggregating cells until sufficient thresholds are reached. These methods have to be applied several times, and this becomes very cumbersome when different geographies are used or when consistency is required between different linked tables. Moreover, post-tabular methods can distort relationships between variables (Kamlet et al. 1985) and spatial correlations.

Pre-tabular methods appear to be a very attractive solution<sup>8</sup>. Firstly, they only have to be applied once, because if micro-data are safe, so all possible aggregations from them will be safe too, and consistency is preserved. Secondly, they are more customisable<sup>9</sup> and they allow a great flexibility of statistical products, both with grid data or hypercubes (they also permit tailored data for users). Another advantage is that some pre-tabular methods (like record swapping) can be unbiased, whereas most post-tabular methods involve suppressing cells and then introducing bias in the estimation of parameters, or turning some parameters not estimable. Nevertheless, in practice, a single table from which every table could be safely extracted, is not realistic, because for a given level of risk, the SDC expert will have to alter too many records (Young et al. 2009), which is not reasonable for an NSI. Moreover, pre-tabular methods can let the users believe that nothing is being done to ensure confidentiality (Longhurst et al. 2007, Shlomo 2007), because applied alone, they can lead to releasing small cells for sensitive variables.

A classic trade-off is to implement basic protection in the micro-data file, and then to add protection to tables when needed (Massell et al. 2006, Hettiarachchi 2013). Post-tabular methods are indeed applied under some conditions for output products (thresholds,  $(n, k)$  rule,  $p\%$  rule, etc.). For example, after perturbation on micro-data, cells that are still unique regarding a given variable will be suppressed.

To take spatial features into account, pre-tabular methods seem to be more appropriate because it is possible to use the geographical information directly to target the riskiest records for the perturbation.

### 14.2.2 Overview of SDC methods taking geography into account

Traditional SDC methods are already the purpose of a dedicated Eurostat handbook (Hundepool et al. 2010, Hundepool et al. 2012) and are therefore not detailed in this chapter. However, we here describe and give references of methods with more explicit consideration of geographic information.

#### Local imputation

Markkula 1999 is one of the first articles that takes the fact of having geographical data in the choice of the SDC method into account. His method, local restricted imputation (LRI), was co-developed by Statistics Finland and the University of Jyvaskyla, and has been tested on census data by Statistics Finland. The method includes three phases:

1. definition of the setting: minimum size of an area, and spatial configuration (3 nested levels called microdata level, macrodata level and stencil level);
  2. identification of risky areas, with the number of individuals under a safety threshold;
  3. imputation of the new values for the risky areas. Two techniques are considered: (i) imputation by the mean of all risky areas belonging to the higher hierarchical level and (ii)
- 
8. The purpose of pre-tabular methods is not to release micro-data themselves, but to provide a common base to build tabular data or grid data.
  9. In general, NSIs do not reveal to the users the setting of parameters of the SDC method (rate of swapped observations, PRAM matrices, parameter of a distribution law, etc.), in order to minimise the risk of retro-engineering by the intruder (Shlomo et al. 2010, Zimmerman et al. 2008).

imputation by random permutation: value in the risky area is replaced by a value from another risky area drawn randomly in the surrounding area.

The LRI method mainly aims at preserving spatial correlations, while restoring as much information as possible about the data. Then it can be appropriate for grid-based data (Tammilehto-Luode 2011). The advantage is that it is simple to understand and offers consistency of results (totals in the higher hierarchical level are preserved), but the method lacks documentation to be precisely reproduced.

### Geographical aggregation

Most of the time, the data protection rule takes the shape of a threshold below which data cannot be disseminated. In the context of grid data, a strategy consists in aggregating contiguous grid cells into bigger polygons (*e.g.* rectangles or bigger squares), so that each polygon respects the threshold. These methods consist in finding the optimal grid where information is released at the most detailed level it can be released. It stands somewhere between SDC methods and data visualisation, because it consists in creating maps of varying resolutions, according to some statistic criteria. New polygons can be obtained by bottom-up aggregation (grouping cells until the threshold is reached), or by disaggregation (starting with a large group of grid cells and cutting it until no sub-cutting is possible any longer).

These methods have interesting properties: for additive data, additivity is preserved, and for average values, the method is perfectly equivalent to using imputation by the mean of the polygon for all squares within a same polygon. This method does not introduce "false zeros" and respects, by construction, the threshold rule.

On the other hand, geographical aggregation does not solve other issues like easy re-identification of extreme values or rare combinations. It also sometimes leads to polygons that do not correspond to any geographical reality: for example, an island can be grouped with the closest mainland. And finally, it also leads to differencing issues with other levels of releasing (co-existence of geometric and administrative boundaries).

Two variations of this principle of geographical aggregation are presented below. The former was developed by INSEE (French National Institute of Statistics) for fiscal data releasing in 2013, and the latter for building stock visualisation in Germany (Behnisch et al. 2013).

**Box 14.2.1 — Exemple 1: a grid of rectangles.** In 2013, to release fiscal data at the level of 200-meter squares - grid-based level in France, INSEE had to respect a regulatory threshold: no fiscal statistic can be released if it does not refer to a minimum of 11 fiscal households. To do this, INSEE used a disaggregation algorithm. The metropolitan territory is first divided into 36 large squares of similar sizes. Each large square is cut horizontally or vertically to form two smaller rectangles. The resulting rectangles are then cut horizontally or vertically, and so on. Horizontal and vertical cuttings always pass through the centre of gravity weighted by the population.

The choice at each stage between horizontal cutting, vertical cutting, or absence of cutting is arbitrated as follows (see Figure 14.1):

- if the horizontal and vertical splits each produce at least one rectangle of less than 11 households then the division is not carried out, in order to respect the constraint of statistical confidentiality
- if only one of the two splits produces a division into two rectangles of 11 or more households each, this split is carried out
- if the two splits each produce two rectangles of more than 11 households, the chosen split is the one that produces two rectangles within which the inhabited squares are least dispersed. The dispersion of a rectangle is measured by the sum of the squared distances

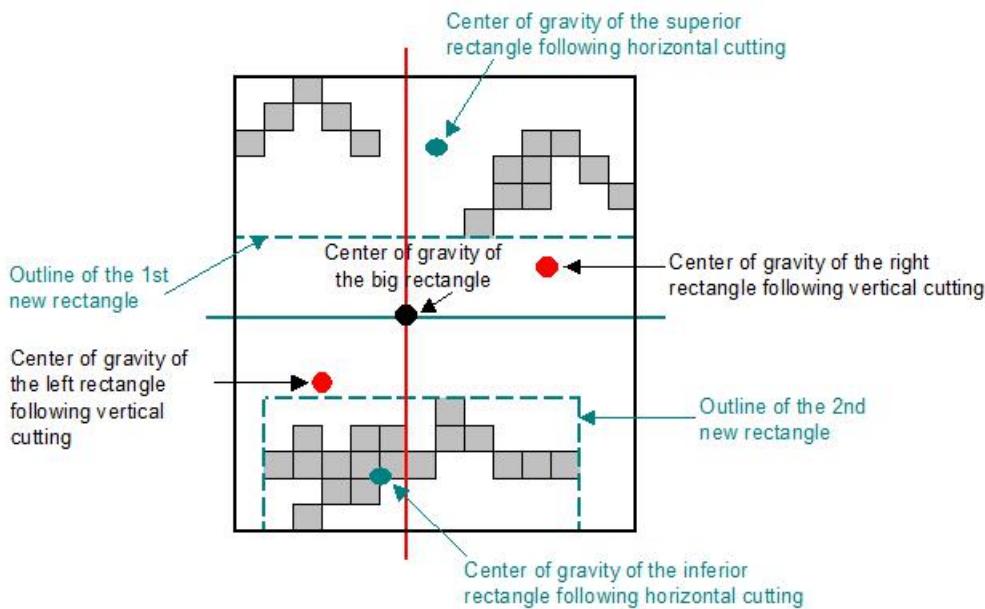


Figure 14.1 – Example of trade-off between horizontal and vertical cuttings

between its centre of gravity and its population-weighted population squares, and the chosen chopping minimizes the sum of the dispersions of the two resulting rectangles. The grid of rectangles reflects well the spatial irregularity of data. However, it is designed conditionally to a version of data set. The grid is not stable for different sources or for different versions of the same source.

**Box 14.2.2 — Exemple 2: Quadtree method.** The quadtree method is another geographical aggregation algorithm that allows multiple resolutions of data in one visualization. It was conducted by the Leibniz Institute of Ecological Urban and Regional Development (Behnisch et al. 2013) for building stock visualization in Germany. The algorithm begins with the highest resolution grid (*e.g.* 250m \* 250m) and if a grid cell contains a number of units under the threshold, it is aggregated with its neighbours to form a cell of bigger level (500m \* 500m). The algorithm stops when all cells are above the threshold (see Figure 14.2).

The quadtree approach offers a consistent grid for different sources, but also for different versions of the same source over the time. It means that it is possible to find a level where different sources can be combined for analysis purposes. The drawback is that it masks some cells above the threshold (in bold on Figure 14.2), and it does not totally solve the MAUP (Modifiable Areal Unit Problem), since data are aggregated into a grid defined in a deterministic way.

### Targeted record swapping

Swapping in general (sometimes considered as a particular case of Post Randomisation Method (PRAM, Gouweleeuw et al. 1998)) consists in exchanging attributes of two observations. Targeted swapping (as opposed to random swapping) targets the riskiest records of the data for exchanging of attributes. This pre-tabular method is often shown as a good compromise between protection and utility. Swapping offers consistency since one record takes the place of another, so that whatever the variables considered, univariate distributions are preserved, and number of records by cell is not

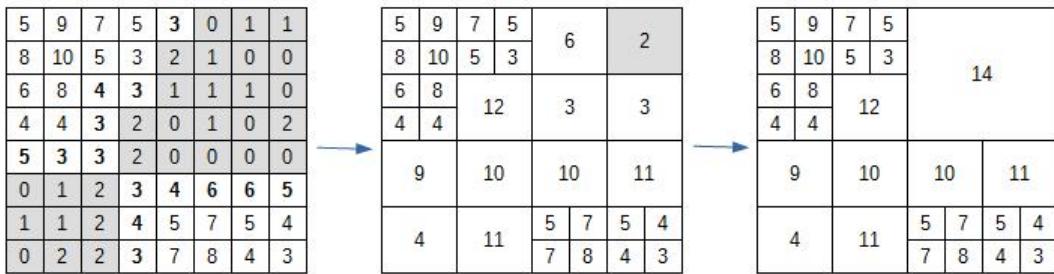


Figure 14.2 – Example of quadtree approach (bottom-up principle) applied to grid data, with a threshold of 3

modified (particularly, swapping does not introduce "wrong zeros" for the global count of records). Inconsistencies only appear when combining variables.

The Office for National Statistics (ONS), the British NSI, had several initiatives about geographically intelligent extensions of swapping for census data releasing (Brown 2003, Shlomo 2007, Young et al. 2009, Shlomo et al. 2010). Targeted record swapping (TRS) is a pre-tabular method that has been tested for some synthetic data from census data in Great Britain. In Japan, Ito et al. 2014 has also tested targeted swapping for the 2005 Census micro-data release. The main addition of TRS is to concentrate the swapping on the observations with the greatest risk of reidentification, defined at the level of a given geography, and to coerce swapped records not to be too distant geographically.

First versions of TRS were developed for hierarchical geographies (Brown 2003, Shlomo 2005), or for grid-based data (Nagy 2015). In these initiatives, two individuals cannot be swapped if they do not belong to the same area at a superior hierarchical level. For a given risky observation, the eligible observations are typically those belonging to the same surrounding area, and among them the match is made on key variables, giving priority to other risky records and eliminating records which have already been swapped.

The local density swapping (LDS) method, described by Young et al. 2009, goes one step further and uses directly the spatial coordinates in the distance function. In LDS, for a given risky observation, the eligible observations are those having the same match variables, and among them a distance function is minimised in order to choose the record to be swapped with. The main idea of LDS is to replace the Euclidean distance by the number of households between the two households to be swapped (*i.e.* located in the circle centred on the original household and with the matching household on the circumference), so as to take population density into account. Priority is given to records which have not already been swapped.

LDS allows a lot of flexibility since it is largely customizable (size of sample, choice of distance, set of matching variables). It also appears to be particularly appropriate to the context of grid data since it takes the geography into account more precisely than the other scenarios. However the drawback of LDS is it is, like every pre-tabular method, not self-sufficient. Moreover the method can leave the impression that nothing has been put in place for disclosure control.

## Extensions

### Trajectories

Trajectory data can be considered as a specific kind of spatial data. Bi-localised data can be collected by a lot of technologies, but they are highly sensitive, because they say a lot about individual habits (places usually visited, etc.). This is why their de-identification is more delicate.

A trajectory is often associated with a temporal aspect, which it is relevant to take into account in the protection method. Both temporal and spatial aspects of the trajectories can be considered when measuring the distance between two trajectories.

Domingo-Ferrer et al. 2011 present two methods to anonymise trajectory data, named *SwapLocations* and *ReachLocations*. The first one preserves trajectory  $k$ -anonymity whereas the second one guarantees location  $l$ -diversity.

Shlomo et al. 2013 suggest a protocol to detect and correct trajectory outliers, taking geographical information into account. The authors consider commuting to work statistics: each trajectory is characterized by two geographical positions and one travel time (in minutes). Outliers are defined according to a given mode of transport. In a first step, outliers are detected among the trajectories, using the Mahalanobis distance (based on a multivariate normal distribution). In a second step, outliers are modified. The place of residence is altered, but the workplace is unchanged so as not to introduce inconsistency (perturbation would be easy to detect if a factory is set where there is none in reality). To do that, the authors define a coherence function at individual level, in order to evaluate the plausibility of the trajectory with respect to the multivariate dataset of non-outliers.

Several algorithms have been tested in this article:

- *record swapping*: iterative algorithm where for each subgroup mode of  $transport \times sex \times age$ , they swap the places of residence while keeping other variables unchanged. At each iteration, consistency is calculated for all possible pairs, and the match is made for records that optimize the consistency. Iterations stop when the general gain of coherence (decreasing at each step) becomes less than a predefined threshold.
- *hot deck*: instead of being exchanged, places of residence for outliers are erased and replaced by imputation from the value taken by a donor having the same characteristics. Selecting the donor can be made by maximising the coherence among all potential donors in a neighbourhood, or minimising the difference in terms of travel time.

Finally, hot deck corrects more outliers than swapping, but swapping minimises the loss of information. In both cases, it is possible that non-outliers become outliers (but fewer cases for swapping).

### **Geomasking**

The term *geomasking* was introduced by Armstrong et al. 1999. It brings together all the methods aiming at altering the geographical position of a point, in order to guarantee more confidentiality for spatial point patterns. One of the most popular geomasking techniques is the donut method, in which every geocoded address is relocated in a random direction, with a distance superior to a minimum and inferior to a maximum.

Geomasking has not been documented much in economics but it is widely used in epidemics or for crime data. In these fields, point patterns have to be released and studied, whereas we assume here that the goal is to release mesh data (on a regular grid or on an administrative zoning), and that micro-data is not the final product but some input we can alter to reach the goal. In the context of census data, moving households are generally excluded, because it could result in obvious inconsistencies (for example it could lead to set a household in the middle of a lake). It also creates "wrong zeros" and does not preserve "true zeros".

#### **14.2.3 How to assess effectiveness of disclosure control?**

##### **Spatial measures of utility**

Applying an SDC method consists in deteriorating data quality in exchange of more protection and results in a loss of information for users. To arbitrate between different SDC scenarios, measuring utility actually means measuring a disutility or a distortion. According to Willenborg et al. 2012 about impact of SDC techniques on micro-data, there are two kinds of losses of information: an increase of the variance in the estimation of a parameter, or the introduction of a bias (which is

obviously the case for example with the suppression of extreme values).

Different metrics can be used to measure the loss of information (Domingo-Ferrer et al. 2001). In all cases, the sharing of perturbed records has to be checked. In addition:

- for continuous variables, mean square error, average absolute error, average rank change, or comparison of Pearson's coefficient between two variables known as correlated. If the dissemination mainly aims at producing a specific indicator like the unemployment rate, it is also relevant to check if bias is not introduced between the original file and the anonymised file. Other model-based metrics can also be used by computing the confidence interval overlap measure for a given logistic regression (De Wolf 2015);
- for categorical variables, it can be direct comparing frequency counts, entropy-based measures like Hellinger distance (Torra et al. 2013), or comparing contingency tables between two variables known as correlated.

For spatial data, we can add to this list the proportion of impacted geographical units, the absolute average deviation (AAD) of countings of a given attribute before and after SDC, calculated at the level of meshes (or squares), or the Moran or LISA indicators<sup>10</sup>, if we know an attribute exhibiting spatial dependency.

### R-U Confidentiality Maps

In order to compare different SDC strategies between them or to choose the most appropriate parameters, "risk-utility maps" can be drawn for different risk levels.

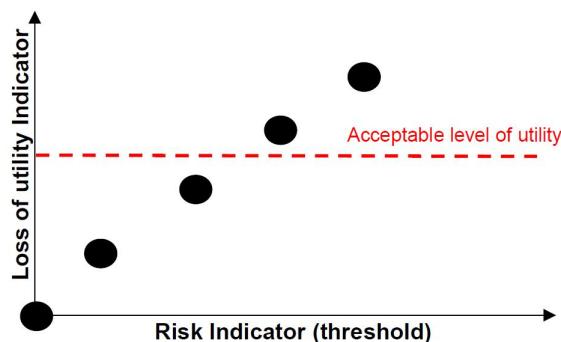


Figure 14.3 – Principle of Risk-Utility Confidentiality Maps

R-U confidentiality maps (Figure 14.3) were first formalised by Duncan et al. 2001 (with an example about additive noise method), but then were used in many papers (Young et al. 2009, Clifton et al. 2012, Gomatam et al. 2005). They constitute a workable tool to frame decision making and to give a synthetic representation of the trade-off between reducing disclosure risk  $R$ , expected low, and preserving data utility  $U$ , expected high. An R-U confidentiality map is a chart that plots the impact in  $R$  and  $U$  of changes in the parameters of a disclosure limitation procedure.

### 14.3 Application for a grid of 1 km<sup>2</sup> squares

In 2017, the Eurostat grant "*Harmonized Protection of Census Data in the ESS*" aimed at harmonising disclosure control techniques concerning Census in European countries, for hypercubes on the one hand and for grid data on the other hand. For this grant, two complementary methods have been chosen, because they seemed to offer a good compromise between confidentiality and utility loss. Targeted record swapping was selected to alter the micro-data in a first step. Then, grid

10. See Chapter 3.

data and hypercubes are built on altered micro-data, and noise is added on cells of hypercubes. This second step is called "Cell-key method" and is inspired from Australian Bureau of Statistics (Fraser et al. 2005).

In this section, we try to assess how targeted record swapping alters spatial correlations, using fiscal data of a small French region. We present the main steps of the method and the results through a risk-utility analysis.

#### 14.3.1 Targeting Record Swapping: details of the method

Implementation choices are taken from an ONS program<sup>11</sup>, and adapted in order to stick to French data. The original algorithm is adapted to hierarchical data, structured in 3 different nested levels ( $level1 \subseteq level2 \subseteq level3$ ). The method is designed in four steps detailed below.

##### Step 1: Targeting the risky records

The first step is to identify the records that need the most to be swapped. An individual can be considered risky or not for a given set of characteristics. Being risky means that there are very few similar records in the same area: a rarity score is computed for each individual as suggested above (average of the reciprocal counts), and individuals with scores above a threshold (a quantile) are flagged as risky. Then, high risk households are defined when there is at least one high-risk individual in the household.

We also associate a geographical level of risk to each individual: if the modality is very rare even at a bigger level (fewer than  $X$  individuals sharing the same modality in the area), then he is "unique" for this geographical level. The geographical level of risk of the household is defined as the highest geographical risk among all persons in the household. A risk 2 household may be matched with a more distant household than a risk 1 household.

##### Step 2: Selection of the sample to swap

The principle of this step is to constitute a sample of risky households, with a size twice smaller than the whole risky population. Then, each household in this sample is associated with another household (preferably risky), so that almost all of the population at risk will be perturbated. We draw a sample stratified by the smallest geographical level, with probability proportional to the arithmetic mean of two indicators (predictors of disclosure):

- a first one increasing with the proportion of high risk households in the geographical unit (this proportion is known throughout the population by construction);
- a second decreasing with the number of households in the geographical unit.

All households have a non-zero chance of being selected for the sample, but high risk households have a much higher probability of being selected. In addition, the sample always contains at least one household per geographical unit. The algorithm also allows to limit the proportion of sampled households in the geographical unit, but results presented below do not use this possibility.

##### Step 3: Matching

The principle of matching is to find, for each household of the sample, a match outside the sample but with close geographical and / or demographical characteristics, with preference for other risky households. The matching process takes place in different stages and sub-stages. Firstly the focus is made for the records risky for the superior level ( $level3$ ), and finally to the inferior level. For each of these 3 stages, constraints of similarity are less and less strict with the sub-stages.

More precisely, if the current stage is dealing with the hierarchical level  $l$ , the principle of each sub-stage is as follows. First, we select a part of the sample. Then, for each household of this sub-sample, we search a "twin" from a "reserve". The match is randomly searched outside the

11. We are grateful to Keith Spicer and Peter Youens for their valuable advice and clarification on the algorithm.

sample. The matching household must have the same profile, be part of a different geographical area (level  $l$ ), but within the same geographical area at the superior hierarchical level (level  $l+1$ )<sup>12</sup>. For example, for a household flagged risky in step 1, with a geographical level of risk  $level1$ , another household will be searched outside the same  $level1$  but inside the same  $level2$ .

There is a preference for households also identified as high-risk. At the end of each stage, the "reserve" is reduced by matched households, so that a household cannot be swapped several times. As long as the sub-stages go on, the constraint on the profile is released, so that at the end all the households in the sample have been matched to another household. This whole method ensures that almost all identified high-risk households are swapped.

#### Step 4: Swapping

Finally, geographical information is swapped. The method does not introduce "false zeros" in counts of people, but it can for counts of specific variables.

### 14.3.2 Choice of data and parameters

#### Data

For this chapter, we chose to apply TRS on exhaustive fiscal data<sup>13</sup>. Unfortunately, tests have only been made on the region Corsica (the smallest French NUTS 2), in order to test many parameters in a reasonable computation time. These tests have been carried out for experimental purposes and should be extended on more populated areas to generalise the conclusions.

For this TRS algorithm, each unit of the 3 needed geographical levels must contain a sufficient number of records. We chose to build *ad hoc* geographical units with INSEE's geographical aggregation algorithm described in section 14.2.2. Squares of 1 km<sup>2</sup> are grouped to constitute rectangles. At the end, we have the following hierarchical structure:

- *level 3*: NUTS 3 (French "départements");
- *level 2*: rectangles containing at least 5000 individuals, intersected with level 3;
- *level 1*: rectangles containing at least 100 individuals, nested in level 2 rectangles (Figure 14.5), and intersected with level 3.

Each level is obtained by disaggregating the previous level, and the most detailed mesh is the 1 km<sup>2</sup> square. If a 1 km<sup>2</sup> square contains at least 100 individuals, then it is not aggregated with neighbour squares to constitute a level 1 unit.

Corsica is made of 2976 1 km<sup>2</sup> squares but only 756 small rectangles (containing at least 100 inhabitants) and 39 big rectangles (containing at least 5,000 inhabitants, see Figure 14.5). Even if rectangles are built to reach a threshold of records (5,000 or 100), all the units do not have the same number of households because some 1 km<sup>2</sup> squares are made of more than 5,000 households in the big cities (Ajaccio or Bastia, Figure 14.4).

In these tests, the geographical level of releasing is not squares but groups of squares (*level1*). If we want then to release counts for quasi-identifiers at a finer level (1 km<sup>2</sup> square), then distribution keys will have to be set, for example randomly in non-empty squares of the *level1* unit, or proportionally to the number of inhabitants of the square if this quantity is known (not sensitive).

#### Parameters

Firstly, we chose 4 categorical variables to define disclosure risk - gender, 5 year age-range, place of birth (12 modalities) and place of residence the previous year (7 modalities).

12. More precisely again, the algorithm proceeds in a set of iterations. At each iteration, each household of the sub-sample is randomly assigned another potential household, and it is decided to match it if all the conditions are met. If this is the case, both households come out of the reserve. Otherwise, the first household remains in the sub-sample and the second one remains in the reserve for the next iteration. A sufficiently large number of iterations is selected so that at the end, no more possible matching can be found.

13. The French Census is specific because it is a survey with associated weights, and this chapter does not aim to discuss the weighting issues.

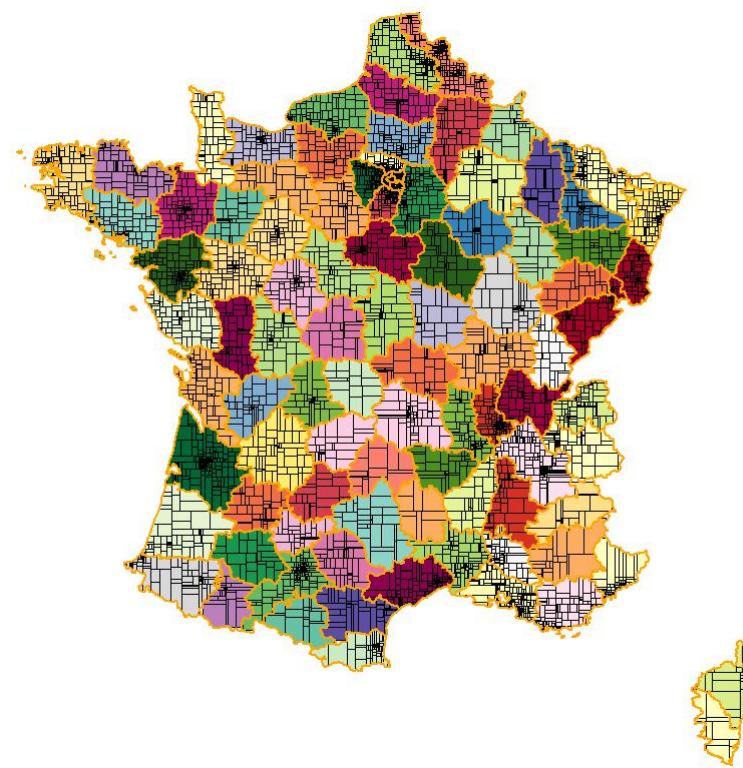


Figure 14.4 – France split into 5,000 individuals rectangles (built at NUTS 3 level)

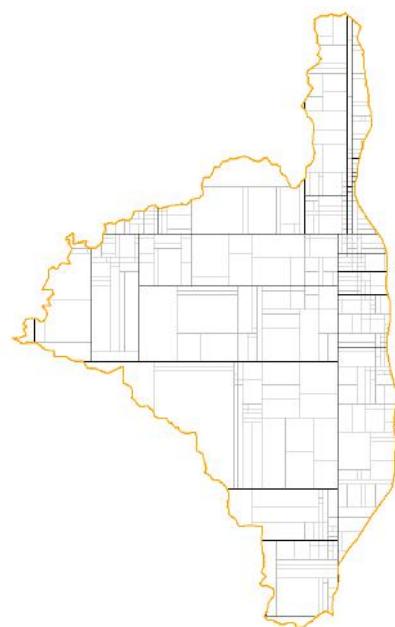


Figure 14.5 – Department 2B (Haute-Corse) split into level 1 and level 2 rectangles

Then, the main parameter is the threshold below which a record will be considered as risky. The sample size and therefore the share of swapped households are derived from it, even though there is no direct formula between the two. By construction, the proportion of swapped households in the population will be slightly higher than this parameter, but of the same order of magnitude. Different parameters (from 1 to 10th percentiles) have been tested, leading to proportions of swapped individuals from 2% to 16%<sup>14</sup>.

Finally, 3 profiles are defined, from the least detailed to the most precise. Two households will not be swapped if they don't share the same profile. For the following simulations, we chose:

- profile A (most detailed): similar number of persons in each of the 7 gender\*age categories<sup>15</sup>;
- profile B (intermediate): similar number of persons in each of 5 gender\*age categories;
- profile C (less detailed): similar number of persons in the household.

### 14.3.3 Results

The output of the algorithm is an altered data set containing, for each record, the original area before swapping, and the area after swapping. Counts can then be made with this output.

Results are shown through a risk-utility analysis (see Section 14.2.3). The risk measure is the threshold set as a parameter of targeted record swapping (from 1 to 10%). A high threshold means that a low level of risk<sup>16</sup> is accepted.

To measure utility loss, the following metrics are used<sup>17</sup>:

- share of level 1 units (small rectangles) impacted by swapping (counts are not the same), for two variables - number of males (taken into account in the matching step) and number of people born in France (not directly taken into account);
- absolute average deviation of countings for level 1 units (small rectangles), for the same two variables;
- Moran's I, calculated at the level of the small rectangles, for 4 sensitive variables with different intensities of spatial autocorrelation: number of people born in France, number of children under 5 years old, income, and number of people belonging to a deprived neighbourhood (QPV<sup>18</sup>).

Results of the tests are shown in Table 14.1 and Figure 14.6 (RU-Maps slightly different from suggested previously). Distortion is measured for variables directly taken into account in the method through the matching profile (V1, number of males), indirectly taken into account in the method through the rarity score (V2, number of people born in France or V3, number of children under 5 years old), or not taken into account at all in the method (V4, income, or V5, number of people in QPV).

We easily see that the higher the acceptable level of risk is (*i.e.* the lower the share of population considered as risky), the lower the distortion in the share of impacted geographical areas and average absolute deviation are.

Even for low values of parameters, the majority of the rectangles are impacted by TRS (for the 1% parameter, the highest level of risk tested, 70% of the counts per small rectangle are changed

---

14. For another region with more inhabitants, the share of swapped records would be closer to the initial parameter. The reason is that in the case of Corsica, the constraint is more difficult to satisfy and the match is more often found outside the risky population.

15. For some age categories both males and females are grouped.

16. We also considered another risk measure with the 90th percentile of the rarity score defined as above (average of the reciprocal counts of the level 1 unit), but this does not vary enough to make relevant graphs.

17. We do not consider the share of swapped individuals for the risk-utility analysis because with this method, it is by construction highly linked to the threshold parameter.

18. In French, "Quartier Politique de la Ville".

Risk measure (%)	0	1	2	3	4	5	7	8	10
Threshold parameter	0	1	2	3	4	5	7	8	10
<b>Utility measures (%)</b>									
Share of swapped individuals	0	2	4	5	7	8	11	13	16
Share of impacted level 1 units - V1	0	38	52	56	63	69	73	75	78
Share of impacted level 1 units - V2	0	71	82	85	85	88	90	92	94
AAD (level1) - V1	0.0	0.5	0.8	0.9	1.1	1.2	1.5	1.6	1.7
AAD (level1) - V2	0.0	0.8	1.1	1.3	1.6	1.6	2.1	2.2	2.5
Moran (level 1): V2	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5
Moran (level 1): V3	6.5	6.4	6.5	6.5	6.6	6.7	6.7	6.8	6.4
Moran (level 1): V4	5.5	5.2	5.1	4.6	5.2	6.8	8.2	5.7	6.4
Moran (level 1): V5	7.7	7.7	7.8	7.7	7.8	7.7	7.3	7.2	7.3

V1: number of males

V2: number of people born in France

V3: number of children under 5 years old

V4: average income

V5: number of people in QPV

Table 14.1 – Results of the tests led on Corsica for several parameters

for the variable "number of people born in France"). But the change is reasonable: for the highest level of risk, 0.4% and 0.7% (respectively for V1 and V2) of the absolute changes are under 5% of the count and the AAD are also under 1%. For the lowest level of risk tested (10% parameter), the AAD is 2.5% for the number of people born in France and 1.7% for the number of males.

Concerning spatial correlations, we now focus on the distortion of Moran's indicator (calculated for *level1* units, before and after TRS). We see that the distortion can be very important (up to 50% of variation of the indicator), that it does not always go in the same direction (the spatial correlation can be increased or decreased with the method), and it is not a monotonic function with level of risk.

Finally, we also see that the utility loss, as regards all the indicators, varies with the variable. If the variable has been taken into account directly in the method (in the definition of the profile: number of males in the tests), then the variable is less distorted than if it has been taken indirectly taken into account (in the identification of high risk people: number of people born in France or number of children under 5 years old in the tests), and *a fortiori* if it has not been taken into account at all (income or belonging to a deprived neighbourhood in the tests).

More specifically about the distortion of spatial correlations: Moran's indicator is unchanged for the variable defining the matching profile (V1), or slightly changed for variables indirectly taken into account (V2 and V3). It is also slightly changed for variables strongly correlated with the matching profile (V5, Figure 14.2). On the opposite, spatial correlations can be very distorted for variables that are not correlated with the matching profile (V4).

The distortion of Moran's I does not particularly increase with the level of risk, but erratic behaviours can appear, due to the randomness of the algorithm during the matching step. Since the method does not consider the income as a variable to preserve, and since this variable is not correlated with another variable that must be preserved, then households with similar incomes can be brought closer or more distant, randomly, from one execution of the method to another.

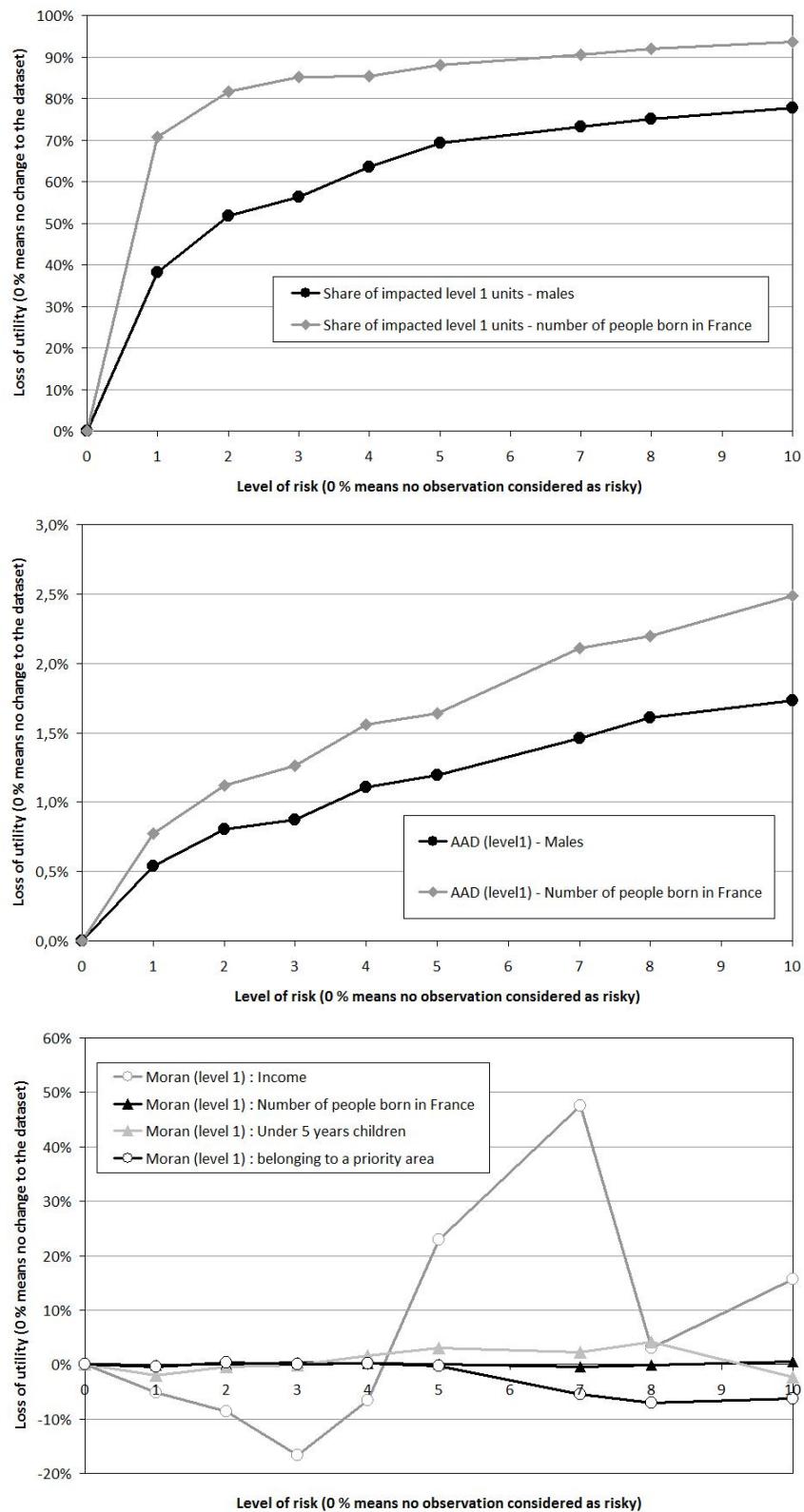


Figure 14.6 – Utility loss as a function of risk level, for 3 utility loss indicators

Pearson coefficient	V1	V2	V3	V4	V5	V6
V1 (number of males)	1	1.00	0.97	0.13	0.97	1.00
V2 (number of people born in France)	1.00	1	0.96	0.14	0.97	1.00
V3 (number of children under 5 years old)	0.97	0.96	1	0.14	0.93	0.97
V4 (average income)	0.13	0.14	0.14	1	0.15	0.13
V5 (number of people in QPV)	0.97	0.97	0.93	0.15	1	0.97
V6 (total number of people)	1.00	1.00	0.97	0.13	0.97	1

Note: V1, V2, V3 and V5 are totals and are strongly correlated to the total number of persons in the rectangle whereas V4 is an average.

Table 14.2 – Pearson coefficients between variables

## 14.4 Differencing issues

### 14.4.1 Definition

Geographic differencing occurs when an intruder can combine data released in various geographies to reconstruct data on a smaller area or deduce the location of an observation. The issue has been presented about Census releasing in many papers (Duke-Williams et al. 1998, ONU 2004), but it occurs for any source releasing.

With nested geographies (*e.g.* regions – departments – towns) the problem is quite simple to solve because the data that can be obtained with subtractions is directly linked to the hierarchy between the various geographies. Thus, once the set of small areas that need to be protected is identified which is called primary secret, SDC software like Tau-Argus can be used to choose the secondary secret. The set of areas that need to be treated so that intruders cannot reconstruct the data of the primary secret. With nested geographies described in a hierarchical tree, the problem is similar to any other variable of interest used in a tabulation (*e.g.* sections – divisions – groups – classes, used in the NACE classification of economic activities).

But the issue of geographic differencing gets more complex when the various geographies used in the release are non-nested (ABS 2015). In that case, there is no hierarchical tree to be used and specific algorithms need to be implemented to identify all the subtractions that an intruder could make between the various areas to get data on smaller areas.

This differencing issue increases when the size of the zone of releasing decreases (blatant in case of small-size grid data). It also increases with the number of various geographies, especially when they are not hierarchical. For example, if NSIs release data on *ad-hoc* zoning in specific partnerships, or if such tailored geographies are constructed by users with web services.

Another example of a differencing issue is when the same phenomenon is observed on various dates. For example, in the case of data about companies released each year, an intruder could compare the various releases to try and find some hidden values. When applying SDC techniques for the latest broadcast, one should therefore take into account what was done for the previous ones and which values were hidden.

### 14.4.2 Illustration

Figure 14.7 presents examples of possible cases of geographical differencing. Overlapping zones between the circles (A) and rectangles (B) are highlighted in orange. In the first case, zoning B encompasses zoning A. An intruder can reconstruct information about B-A by subtraction and

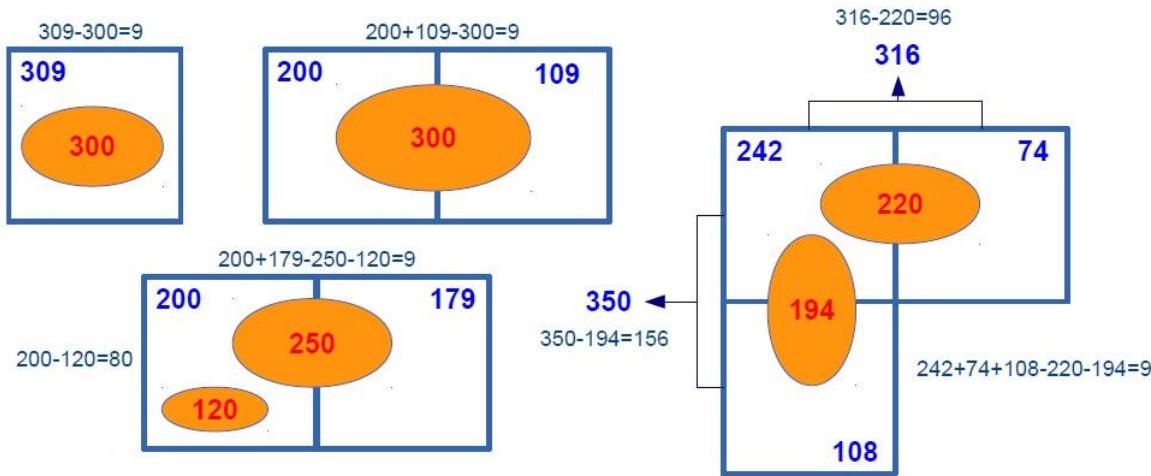


Figure 14.7 – Confidentiality breach by geographic differencing

this may lead to revealing data concerning a small number of individuals. In the second box, the intruder can combine two zones of zoning B to perform the operation  $(B_1 + B_2) - A$  and thus obtain information concerning a non-released zone. The last two cases show that differencing can occur with a combination of any number of the two zonings.

With the frequency counts included in these examples, if information cannot be disseminated if it concerns less than 10 individuals, then there is a breach of confidentiality by geographical differentiation in each of these four examples. In other cases of overlapping, the intruder cannot directly obtain information on a new zone, but by taking into account auxiliary information or the geography surrounding the overlapping area, problems may arise. The geography of the zone needs to be taken into account as it is sometimes impossible for certain areas to contain any observation (lake, highway, etc.). These empty areas cannot be used to protect the data and must be disseminated.

#### 14.4.3 Identifying Risky Areas

The first step in order to deal with the differencing issue is to flag the risky areas. This can be relatively simple with the geographic information system (GIS), but it becomes complicated when the number of non-nested geographies increases because it increases the dimension of the problem to solve and can lead to NP-hard problem.

If the choice is to suppress the information for these risky areas, the method is carried in two steps: primary then secondary suppressions.

The algorithm needs to look for the possible overlaps between the non-nested geographies. As seen in Figure 14.7, problems can arise with a combination of multiple areas of each zoning. A confidentiality criteria needs to be chosen, for example at least 10 individuals in any area.

The algorithm needs to include checks with the totals if one of the non-nested geographies is hierarchical (for example when on the one hand data is released for regions-department-town but on the other hand data is released for a partner with a specific zoning).

It might be important to reduce the information loss and thus to include optimization rules to minimize the number of individuals impacted, or give priority to keep preserved areas if information is more useful there, for example when conceiving public policies for deprived neighbourhoods. This work requires a consequent disk space and a lot of computing power: the algorithm might need a long time to explore all the possible overlaps.

#### 14.4.4 Protection methods

Different types of methods can be used to restore confidentiality when facing differencing issues due to such overlaps.

First, the zoning can be modified: the boundaries can be changed to eliminate areas of overlap, for example by nesting the various geographies and creating a clean hierarchical tree.

Secondly, if boundaries are fixed, various zones can be merged to eliminate overlaps. It reduces the levels of detail but it enables the data provider to release information on these areas.

A third method is to suppress data for specific areas where overlapping occurs. Due to the constraints in the data production system, this option is often chosen when confidentiality problems arise and it leads to a trade-off between the level of details of the release and the number of hidden areas due to differencing issues.

Instead of suppressing data when boundaries and zoning are fixed, data can be perturbed for example by adding or subtracting small numbers to the risky areas in case of frequency counts release. To do this in a consistent way for multiple tables, or for various geographies, ABS has conceived a cell key method that makes use of record keys assigned to each micro-data observation to keep the perturbation consistent between the various geographies (Fraser et al. 2005). The cell key method was adapted by ONS for the census release and the method was also tested for the Eurostat "Harmonized Protection of Census Data in the ESS" Grant.

### Conclusion

Reflection about SDC methods goes hand in hand with a strategic reflection of the NSIs on what they want to release *in fine*. Aversion to the dissemination of "false" information and fear of misreading of hurried users must be discussed. Choices also have to be made whenever possible in consultation with potential future users, which is the best way to preserve statistical relationships that will be analysed at the end.

Dealing with spatial data can be seen as an opportunity to refine SDC methods, because density and dissimilarity with the neighbours are a fundamental predictor of disclosure risk. In the actual state of the art, geographical information is taken into account by perturbing the micro-data using local information of the neighbourhood (local imputation, targeted record swapping).

In the future, in conjunction with increasing computing capacities, the geographical coordinates may be used more precisely, for example by measuring local density for each record. But precision improvement must be balanced against the extra-complexity of the SDC method and the inherent additional difficulty to communicate to the users about the protection method.

Tests led on exhaustive fiscal data for one region of France show that for reasonable risk levels, targeted record swapping implies low distortion of spatial correlations, even if these tests would deserve to be continued with other SDC strategies and on bigger regions.

Nevertheless, pre-tabular SDC method is not sufficient by itself, firstly because reaching an acceptable level of global risk in the data set would require perturbing too many records, and secondly because of public perception. Post-tabular methods make more visible the existence of disclosure protection to respect the regulatory threshold. This is why concerning the census, the advice given by Eurostat to SDC experts is to combine pre-tabular methods taking geographical specificities into account, and post-tabular methods.

In any case, whatever the method used, and even if it is the most traditional, it is interesting to measure how much the SDC technique degrades the spatial relationships of some attributes. For this purpose, R-U confidentiality maps drawn for distortion of spatial correlations coefficients make an efficient operational tool.

Although precise parameters used might be kept secret to improve the protection, it is necessary that NSIs and data providers document the method and the choices that were made. Users must be

conscious that the data has been changed or might be incomplete when conducting their analysis. For example, the SDC expert can then communicate to potential users how much Moran's I or LISA are affected, in order to guard users against any misleading use of the protected data.

## References - Chapter 14

- ABS (2015). « SSF Guidance Material – Protecting Privacy for Geospatially Enabled Statistics: Geographic Differencing ».
- Armstrong, Marc P, Gerard Rushton, Dale L Zimmerman, et al. (1999). « Geographically masking health data to preserve confidentiality ». *Statistics in medicine* 18.5, pp. 497–525.
- Backer, Lars H et al. (2011). « GEOSTAT 1A: Representing Census data in a European population grid ». *Final Report*.
- Behnisch, Martin et al. (2013). « Using Quadtree representations in building stock visualization and analysis ». *Erdkunde*, pp. 151–166.
- Bergeat, Maxime (2016). « La gestion de la confidentialité pour les données individuelles ». *Document de travail INSEE M2016/07*.
- Brown, D (2003). « Different approaches to disclosure control problems associated with geography ». *Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*.
- Clarke, John (1995). « Population and the environment: complex interrelationships. »
- Clifton, Kelly and Nebahat Noyan (2012). « Framework for Applying Data Masking and Geo-Perturbation Methods to Household Travel Survey Datasets ». *91st Annual Meeting of Transportation Research Board, Washington, DC*.
- Curtis, Andrew J, Jacqueline W Mills, and Michael Leitner (2006). « Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina ». *International Journal of Health Geographics* 5.1, pp. 44–55.
- De Wolf, PP (2015). « Public use files of eu-silc and eu-lfs data ». *Joint UNECE-Eurostat work session on statistical data confidentiality, Helsinki, Finland*.
- Deichmann, Uwe, Deborah Balk, and Greg Yetman (2001). « Transforming population data for interdisciplinary usages: from census to grid ». *Washington (DC): Center for International Earth Science Information Network* 200.1.
- Domingo-Ferrer, Josep, Josep M Mateo-Sanz, and Vicenç Torra (2001). « Comparing SDC methods for microdata on the basis of information loss and disclosure risk ». *Pre-proceedings of ETK-NTTS*. Vol. 2, pp. 807–826.
- Domingo-Ferrer, Josep and Rolando Trujillo-Rasua (2011). « Anonymization of trajectory data ».
- Doyle, Pat et al. (2001). « Confidentiality, disclosure, and data acces: theory and practical applications for statistical agencies ».
- Duke-Williams, Oliver and Philip Rees (1998). « Can Census Offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure ». *International Journal of Geographical Information Science* 12.6, pp. 579–605.
- Duncan, George T, Sallie A Keller-McNulty, and S Lynne Stokes (2001). « Disclosure risk vs. data utility: The RU confidentiality map ». *Chance*. Citeseer.
- Duncan, George T and Diane Lambert (1986). « Disclosure-limited data dissemination ». *Journal of the American statistical association* 81.393, pp. 10–18.
- Elliot, Mark J et al. (2005). « SUDA: A program for detecting special uniques ». *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, pp. 353–362.
- Elliot, Mark and Josep Domingo-Ferrer (2014). « EUL to OGD: A simulated attack on two social survey datasets ». *Privacy in Statistical Databases*. Ed. by Josep Domingo-Ferrer.
- Fraser, Bruce and Janice Wootton (2005). « A proposed method for confidentialising tabular output to protect against differencing ». *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, pp. 299–302.
- Gomatam, Shanti et al. (2005). « Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers ». *Statistical Science*, pp. 163–177.

- Gouweleeuw, JM, Peter Kooiman, and PP De Wolf (1998). « Post randomisation for statistical disclosure control: Theory and implementation ». *Journal of official Statistics* 14.4, pp. 463–478.
- Haldorson, Marie et al. (2017). « A Point-based Foundation for Statistics: Final report from the GEOSTAT 2 project ». *Final Report*.
- Hettiarachchi, Raja (2013). « Data confidentiality, residual disclosure and risk mitigation ». Working Paper for joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.
- Hundepool, Anco et al. (2010). « Handbook on statistical disclosure control ». *ESSnet on Statistical Disclosure Control*.
- Hundepool, Anco et al. (2012). « Statistical disclosure control ».
- Insee (2010). « Guide du secret statistique ». *Documentation INSEE*.
- Ito, Shinsuke and Naomi Hoshino (2014). « Data swapping as a more efficient tool to create anonymized census microdata in Japan ». *Privacy in Statistical Databases*, pp. 1–14.
- Kamlet, MS, S Klepper, and RG Frank (1985). « Mixing micro and macro data: Statistical issues and implication for data collection and reporting ». *Proceedings of the 1985 Public Health Conference on Records and Statistics*.
- Lambert, Diane (1993). « Measures of disclosure risk and harm ». *Journal of Official Statistics* 9.2, pp. 313–331.
- Longhurst, Jane et al. (2007). « Statistical disclosure control for the 2011 UK census ». *Joint UNECE/Eurostat conference on Statistical Disclosure Control, Manchester*, pp. 17–19.
- Markkula, Jouni (1999). « Statistical disclosure control of small area statistics using local restricted imputation ». *Bulletin of the International Statistical Institute (52nd Session)*, pp. 267–268.
- Massell, Paul, Laura Zayatz, and Jeremy Funk (2006). « Protecting the confidentiality of survey tabular data by adding noise to the underlying microdata: Application to the commodity flow survey ». *Privacy in Statistical Databases*. Springer, pp. 304–317.
- Nagy, Beata (2015). « Targeted record swapping on grid-based statistics in Hungary ». *Submission for the 2015 IAOS Prize for Young Statisticians*.
- ONS (2006). « Review of the Dissemination of Health Statistics: Confidentiality Guidance ». *Working Paper 5: References and other Guidance*.
- ONU (2004). « Manuel des systèmes d’information géographique et de cartographie numérique ». F-79, pp. 118–119.
- Shlomo, Natalie (2005). « Assessment of statistical disclosure control methods for the 2001 UK Census ». *Monographs of official statistics*, pp. 141–152.
- (2007). « Statistical disclosure control methods for census frequency tables ». *International Statistical Review* 75.2, pp. 199–217.
- Shlomo, Natalie and Jordi Marés (2013). « Comparison of Perturbation Approaches for Spatial Outliers in Microdata ». *the Cathie March Centre for Census and Survey Research*.
- Shlomo, Natalie, Caroline Tudor, and Paul Groom (2010). « Data Swapping for Protecting Census Tables ». *Privacy in statistical databases*. Springer, pp. 41–51.
- Tammilehto-Luode, Marja (2011). « Opportunities and challenges of grid-based statistics ». *World Statistics Congress of the International Statistical Institute*.
- Torra, Vicenc and Michael Carlson (2013). « On the Hellinger distance for measuring information loss in microdata ». *Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada, 28-30 October 2013*.
- VanWey, Leah K et al. (2005). « Confidentiality and spatially explicit data: Concerns and challenges ». *Proceedings of the National Academy of Sciences* 102.43, pp. 15337–15342.
- Willenborg, Leon and Ton De Waal (2012). *Elements of statistical disclosure control*. Vol. 155. Springer Science & Business Media.

- Young, Caroline, David Martin, and Chris Skinner (2009). « Geographically intelligent disclosure control for flexible aggregation of census data ». *International Journal of Geographical Information Science* 23.4, pp. 457–482.
- Zimmerman, Dale L and Claire Pavlik (2008). « Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data ». *Geographical Analysis* 40.1, pp. 52–76.



# Index

- absolute measure, 94  
aggregate, 82  
aggregated distribution, 82  
Akaike criterion - AIC, 240  
Analytical test, 90  
areal data, 6, 23  
bandwidth, 206, 211, 223, 238  
benchmarking, 320  
Best Linear Unbiased Predictor - BLUP, 313  
block kriging, 136  
bottom-up selection, 156  
cartographic semiology, 7  
centrality, 335  
centroid, 23  
change of support, 134  
choropleth maps, 8, 19  
class analysis, 8  
cliques, 338  
cokriging, 136  
communities, 337  
Complete Spatial Randomness - CSR, 76  
completely random distribution, 81  
confidentiality, 212  
continuous data, 5  
Cross Validation criterion - CV, 240  
data – Beischmiedia pendula, 107  
data – Murchison, 107  
data – paracou16, 89, 99, 100, 103  
Delaunay, 34  
Delaunay triangulation, 25  
ecological errors, 162, 286  
economic competition, 151  
edge effect, 76  
edge effects, 206, 210  
Empirical Best Linear Unbiased Predictor - EBLUP, 315  
empirical dispersion variation, 134  
ESPON, 27  
estimator – composite, 314  
estimator – direct, 311  
estimator – Henderson, 313  
estimator – Horvitz-Thompson, 248, 257  
estimator – mixed, 314  
estimator – pseudo direct, 314  
estimator – quality, 316  
estimator – synthetic, 314  
feedback effect, 157, 185  
first-order property of a point process, 78  
flow, 328  
forecast, 313  
free scale networks, 331  
function – autocorrelation, 116  
function – Baddeley, Møller, Waagepetersen’s  $K_{inhom}$ , 88  
function – Besag’s  $L$ , 86  
function – covariance, 116  
function – Diggle and Chetwynd’s  $D$ , 87  
function – Duranton and Overman’s  $K_d$ , 95  
function – intertype, 102  
function – intertype  $K$ , 102  
function – intertype  $M$ , 103  
function – Marcon and Puech’s  $M$ , 96  
function – random, 114  
function – Ripley’s  $K$ , 83  
Geary index, 56  
generalised least square, 313  
geographically weighted regression, 171  
Geostat 2, 256  
geostatistics, 114  
Getis and Ord index, 62  
graph – k-regular, 328  
graph – neighbourhood, 34  
graph – random, 328  
graph – small world, 328  
graph – theory, 333  
graphs, 327  
graphs partitioning, 327  
GRTS, 41, 262  
GRTS ordering, 266  
Hamilton path, 258  
heteroskedasticity, 161  
homogeneity, 77  
homogeneous Poisson process, 76, 77  
imputation, 290  
inclusion probabilities, 257  
independence, 77

- inhomogeneous Poisson process, 79  
 intensity, 78  
 intrinsic stationarity, 115  
 isotropy, 80
- join count statistics, 60
- kernel, 206, 209, 236  
 knn, 37  
 kriging, 5, 127  
 kriging variance, 128
- LISA, 62
- local spatial autocorrelation indices, 6
- Mapinfo, 16
- maps in proportional symbols, 18
- marked point process, 75
- matrix standardisation, 44
- MAUP, 57, 74, 93, 134, 162, 206
- maximum likelihood, 315
- mean squared error, 320
- method – classes of the same amplitude, 8
- method – cube, 263
- method – division, 341
- method – Jenks, 8
- method – k-means, 8, 260
- method – quantile, 8
- method – spatial cube, 264
- MINIMAX property, 257
- model – cardinal sine, 124
- model – exponential, 124
- model – Fay and Herriot, 311, 315
- model – fixed effects, 181, 186
- model – gaussian, 124
- model – generalised linear mixed, 310, 312
- model – hedonic, 232
- model – linear mixed, 311
- model – logistic, 310
- model – Manski, 154
- model – Matern, 124
- model – mixed linear, 307
- model – Poisson, 310, 312
- model – pooled data, 181
- model – power, 124
- model – random effects, 181, 187
- model – RE-SEM (KKP), 184
- model – SME-RE, 184
- model – Spatial Auto Regression (SAR), 155, 183, 280
- model – Spatial Autoregressive Confused (SAC), 155
- model – Spatial AutoRegressive with Autoregressive Disturbance (SARAR), 183
- model – Spatial Durbin (SDM), 155, 183
- model – Spatial Durbin Error (SDEM), 155, 183
- model – spatial dynamic, 198
- model – Spatial Error (SEM), 155, 183
- model – Spatial Lag X (SLX), 155
- model – spatial multidimensional, 199
- model – spherical, 123
- model with variable coefficients, 234
- modularity, 338
- Monte Carlo method, 88
- Moran index, 56
- Moran indicator, 153
- Moran’s diagram, 53
- multi-type process, 98
- multicollinearity, 250
- multiple testing, 65, 252
- neighbourhood matrix, 32
- neighbourhood of points, 75
- neighbourhood Queen Rook, 39
- nonparametric regression, 215
- normality hypothesis, 55
- observation window, 75
- package – BalancedSampling, 263
- package – btb, 217
- package – cartography, 12
- package – dbmss, 89, 91, 97, 99–103, 105–107
- package – deldir, 24
- package – dplyr, 23
- package – ETAS, 83
- package – geoR, 118
- package – gstat, 266
- package – GW.model, 240
- package – igraph, 330
- package – leaflet, 20
- package – maps, 191
- package – maptools, 191
- package – plm, 191
- package – rgdal, 12
- package – rgeos, 12
- package – RgoogleMaps, 20
- package – sae, 321
- package – sf, 20

- package – *sp*, 12, 137, 191  
package – *spaMM*, 321  
package – *spatstat*, 73, 77–82, 84–89, 91, 97, 99–102, 105, 107, 217  
package – *spdep*, 34, 59  
package – *splm*, 191  
package – *stargazer*, 193  
package – *TSP*, 41  
panel model with common factor, 200  
peer effects, 154  
Permanent database of facilities - BPE, 73, 91, 97, 101, 106  
point data, 4, 73, 83, 99  
Primary Sampling Units, 257  
process intensity, 206  
projection system, 18  
proportional symbols maps, 10  
quantile smoothing, 217  
random point process, 74  
randomisation hypothesis, 55  
regionalized variable, 114  
regular distribution, 81  
regularised variable, 134  
relative measure, 94  
ridge regression, 252  
Robust Geographically Weighted Regression, 240  
sampling, 256, 306  
sampling – bias, 306  
sampling – areolar, 283  
sampling – Bernoulli, 282  
sampling – cluster, 283  
sampling – Correlated Poisson, 262  
sampling – determinantal, 267  
sampling – ordered, 265  
sampling – pivotal, 263  
sampling – Poisson, 282  
sampling – simple random, 282  
sampling – spatial pivotal, 263  
sampling – Spatially Correlated Poisson, 262  
sampling – stratified, 284  
sampling – variance, 306  
sampling design, 256, 306  
second order stationarity, 115  
second-order properties of point patterns, 79  
shapefile, 16  
shortest path, 40, 258  
shrinkage, 319  
Simultaneous AutoRegressive model – SAR, 309  
size effect, 285  
small areas, 305  
smoothing, 205  
solid colour maps, 8, 19  
spatial autocorrelation, 52, 152  
spatial autocorrelation indices, 55, 153, 164  
spatial autocorrelation of categorical variables, 60  
spatial data confidentiality, 349  
spatial dependence, 52, 53, 152  
spatial distribution of points, 71, 75  
spatial econometric on panel data, 179  
spatial econometrics, 149  
spatial econometrics of survey data, 277  
spatial heterogeneity, 152, 161, 232  
spatial prediction, 247  
spatial relationships, 32  
spatial sampling, 255  
spatio-temporal indices, 68  
specification tests for spatial econometric, 155  
specification tests for spatial econometric on panel data model, 189  
stationary process, 80  
strict stationarity, 115  
support, 134  
survey, 255, 256, 282, 305  
test stationarity coefficients, 247  
top-down selection, 156  
topographical measure, 94  
trajectory optimisation, 258  
universal kriging, 139  
variogram, 117  
variogram adjustement, 125  
variogram cloud, 119  
Verdoorn’s law, 190  
Voronoi Index, 269  
voronoi, 23  
voronoi polygons, 23  
weak stationarity, 115  
weight matrix, 42, 153, 160, 236  
weighted least squares, 235  
window, 75



The United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM) “acknowledged the critical importance of integrating geospatial information with statistics and socio-economic data and the development of a geospatial statistical framework”. There is a growing importance of taking spatial phenomena into account and, as a consequence, the need for geolocated data are also increasing. To this end, the French national institute of statistics and economic studies (Insee) has undertaken a reworking of its geographical information system. This handbook, coordinated by Insee thanks to a Eurostat grant, shows what kind of analysis can be achieved and the pitfalls one has to avoid once geolocated data is available.

The purpose of this spatial analysis handbook is to answer the questions faced by research teams at statistical institutes. What use should be made of the new geolocated data sources? In what cases should their spatial dimension be taken into account? How should spatial statistical and spatial econometric methods be applied? In contrast to existing manuals, the teaching principle of this handbook has been designed expressly according to the issues specific to statistical institutes, such as spatial sampling, spatial econometrics, confidentiality or spatial smoothing.

The handbook is divided into four parts. The first three match the stages one would follow to carry out a study: describing the location of the observations, measuring spatial interactions and applying the appropriate model. Each of the fourteen chapters deals with a specific subject by explaining the theoretical foundations, giving educational examples based on data coming from public statistical institutes, and displaying how to use the R statistical software to carry out the computations.