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.