Spatial Econometric Methods in Agricultural Economics Using R

Editors

Paolo Postiglione
Professor in Economic Statistics, Department of Economic Studies
University “G. d’Annunzio” of Chieti-Pescara, Pescara, Italy

Roberto Benedetti
Professor in Economic Statistics, Department of Economic Studies
University “G. d’Annunzio” of Chieti-Pescara, Pescara, Italy

Federica Piersimoni
Senior Researcher
Italian National Institute of Statistics, Processes Design and Frames Service, Rome, Italy
Preface

The collection of spatial agricultural data and the spatial analysis of agriculture represent two issues of primary relevance for a large number of people. This book aims at supporting stakeholders to design spatial surveys for agricultural data and/or to analyse the geographically collected data. Hence, the book represents a comprehensive guide in methodological and empirical advanced techniques for practitioners.

This volume can also be considered as a primary tool for users from less developed countries, where agriculture is still the prevalent economic sector. Therefore, different contributions may guide one through the application of spatial survey methods, technologies developed in the past decades, such as remote sensing and GIS, and appropriate methods to analyse spatial agricultural data. Applied spatial analysts might also benefit from this work. In particular, a part of the book is devoted to the integration techniques used to merge agricultural data from different sources. Finally, both people from Academic institutions and National Statistical Offices may appreciate the occasion of deepening their knowledge of spatial techniques for agriculture.

Although the book could also represent a valued support on spatial methodologies in agriculture for graduate classes, the primary audience is mainly composed by researchers with some prior background in econometrics and spatial statistics.

The main objective of this book is to introduce agricultural economists to statistical approaches for the analysis of spatial data. The aim is to illustrate, for the main typologies of agricultural data, the most appropriate methods for the analysis, together with a description of available data sources and collection methods.

Spatial econometrics methods for different types of data are described and adopted with reference to typical analyses of agricultural economics. Topics such as spatial interpolation, point patterns, spatial autocorrelation, survey data analysis, small area estimation, regional data modelling, and spatial econometrics techniques are covered jointly with issues arising from the integration of several data types. Besides, the different phases of agricultural data collection, analysis, and integration are described in a simple way. The joint use of statistical methods, new technologies, and economic theory is treated considering the peculiarities of spatial data for a proper and efficient analysis of agricultural data.

Theoretical aspects of each model are described and complemented by examples on real data that are developed by using the open-source R software. The codes are available in the text, explained with details and in an intuitive way so that the readers
can replicate these analyses on their own data. Moreover, any prior knowledge of the
R programming environment is not assumed throughout the book.

The volume is organized in a number of review chapters on several specific
themes. In particular, this book contains 13 Chapters, of which the first one can be
considered as an introductory chapter, reviewing the main underlying concepts and
presenting each contribution.

We would like to thank Alfredo Cartone for reading some parts of this book and
for his support in the implementation of some R codes. Thanks also to Vijay Primlani
of Science Publishers, CRC Press, for his continuous encouragement to complete
this book. Finally, we are grateful to the individual chapter authors for their diligence
in writing the documents. We are confident that their work will lead to new insights
in the application of spatial econometric methods to agricultural data.

Rome, December 24th, 2020

Paolo Postiglione
Roberto Benedetti
Federica Piersimoni
## Contents

**Preface**

1. Basic Concepts  
   *Paolo Postiglione, Roberto Benedetti and Federica Piersimoni*

2. Spatial Sampling Designs  
   *Francesco Pantalone and Roberto Benedetti*

3. Including Spatial Information in Estimation from Complex Survey Data  
   *Francesco Pantalone and Maria Giovanna Ranalli*

4. Yield Prediction in Agriculture: A Comparison Between Regression Kriging and Random Forest  
   *Eugenia Nissi and Annalina Sarra*

5. Land Cover/Use Analysis and Modelling  
   *Elisabetta Carfagna and Gianrico Di Fonzo*

6. Statistical Systems in Agriculture  
   *Cecilia Manzi and Federica Piersimoni*

7. Exploring Spatial Point Patterns in Agriculture  
   *M Simona Andreano and Andrea Mazzitelli*

8. Spatial Analysis of Farm Data  
   *Alfredo Cartone and Domenica Panzera*

9. Spatial Econometric Modelling of Farm Data  
   *Anna Gloria Billè, Cristina Salvioni and Francesco Vidoli*

10. Areal Interpolation Methods: The Bayesian Interpolation Method  
    *Domenica Panzera*
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.</td>
<td>Small Area Estimation of Agricultural Data</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td>Gaia Bertarelli, Francesco Schirripa Spagnolo, Nicola Salvati and Monica Pratesi</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>Cross-sectional Spatial Regression Models for Measuring Agricultural $\beta$-convergence</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>Alfredo Cartone and Paolo Postiglione</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>Spatial Panel Regression Models in Agriculture</td>
<td>254</td>
</tr>
<tr>
<td></td>
<td>Paolo Postiglione</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Index</td>
<td>275</td>
</tr>
<tr>
<td></td>
<td>About the Editors</td>
<td>279</td>
</tr>
</tbody>
</table>
Chapter 1
Basic Concepts

Paolo Postiglione,1,* Roberto Benedetti1 and Federica Piersimoni2

1.1 Introducing space in agricultural economics

There is a great need for reliable data in agriculture. However, this information should be analysed through appropriate statistical methods to obtain evidence that can assist farmers, for example, to optimize farm returns, decrease unnecessary applications of fertilizers and pesticides, and preserve natural resources.

Standard statistical techniques often perform poorly when employed to agricultural data, due to its spatial nature. In fact, one of the common assumptions in traditional statistics is the independence and homogeneity among the observations, a hypothesis patently violated when applied to agricultural data. An agricultural variable could often display similar values in adjacent areas, leading to spatial clusters. In many cases, nearby fields have similar soil-type, climate, precipitation or an area cultivated with wheat may be close to other wheat-cultivated zones. Ignoring this dependence when analysing agricultural data may produce bias or inefficient estimates. For these reasons, the topic of statistical analysis of spatial data is worth a specific treatment.

For a long time, the analysis of geographically distributed phenomena in agricultural economics was carried out without the consideration of space as crucial information. The present book tries to fill this gap by highlighting potential applications of spatial analysis for agricultural facts. Indeed, space is very important in agricultural economics studies. The land is a crucial resource in agriculture and most of the data collected is spatially distributed. Besides, all agricultural activities are spatially located. However, the application of spatial models has grown to become important in applied agricultural economics only during the last few decades (Anselin 2002; Anselin and Bera 1998; Goodchild et al. 2000).

1 “G. d’Annunzio” University of Chieti-Pescara, Department of Economic Studies.
   Emails: benedett@unich.it; piersimo@istat.it
* Corresponding author: postigli@unich.it
Generally speaking, the term spatial means that each unit has a geographical reference, i.e., we know where each case happens on a map. If the locations of these sites are observed and enclosed to the observations as labels, the resulting data is called spatial data. In spatial data analysis, the set of spatial locations are considered as essential information in the study. Our main idea is that location matters in agriculture and that the occurrences essentially follow the First Law of Geography, which according to Tobler (1970) states that: “everything is related to everything else, but near things are more related than distant things”.

A proper definition of new spatial data science methods is required in order to analyse agricultural data and to uncover interesting, useful, and non-trivial patterns. The first economist that explicitly claimed the importance of space in agricultural economics was von Thünen (1783–1850). Von Thünen developed the Isolated State model (von Thünen 1826), whose framework is considered to be the first serious application of spatial economics and economic geography in agriculture. His agricultural location theory conjectured that the optimal organisation of agricultural activities is based on location factors. Hence, these activities are arranged in concentric rings around a central consumers’ town.

Recently, in many countries, the National Statistical Office geo-references the sampling frames of physical or administrative bodies used in agricultural surveys, not only with reference to the codes of a geographical taxonomy, but also adding data regarding the spatial position of each record.

Modern tools, such as GIS and remote sensing, are increasingly used in the monitoring of agricultural resources. For example, the developments in GIS technology offer growing opportunities to agricultural economics analysts dealing with large and detailed spatial databases, allowing them to combine spatial information from different sources and to produce different models, as well as tabular and graphic outputs.

The availability of these valuable sources of information makes the advanced models suggested in the spatial statistic and econometric literature applicable to agricultural economics.

More formally, spatial statistics is a field of spatial data analysis in which the observations are modelled using random variables. Ripley (1981) defines spatial statistics as “the reduction of spatial patterns to a few clear and useful summaries”, comparing such statistics “with what might be expected from theories of how the pattern might have originated and developed”.

Conversely, spatial econometrics is a branch of econometrics that deals with the modelling of spatial interaction and spatial heterogeneity in data analysis. The birth of this discipline can be traced back to the works of Paelinck and Klaassen (1979) and Anselin (1988). Following Anselin (1988), spatial econometrics can be defined as: “the collection of techniques that deal with the peculiarities caused by space in the statistical analysis of regional science methods”. In this book, we use a broad definition, referring to the term regional as those spatial units defined as areal regions, locations (i.e., points), and continuous units. Essentially, spatial econometrics represents a toolkit that allows for the rigorous treatment of data that is geographically distributed.
Interestingly, the analysis of agricultural yield data was also the motivation for the seminal paper by Whittle (1954), in which he analysed such data through two-dimensional stochastic models.

This book contains several contributions focused on spatial data and its use in monitoring agricultural resources, farms management, and regional markets. The theory of spatial methods is complemented by real and/or simulated examples implemented through the open-source software R.

The layout of this introductory chapter is as follows: in Section 1.2 the main typologies of spatial data are described. Section 1.3 contains some brief considerations about the R software. Section 1.4 outlines the contributions of this book, stressing the main evidences.

1.2 Spatial concepts: the essential

1.2.1 Spatial effects

Spatial econometrics aims to address in a formal way two effects that are typical of geo-referenced data: spatial dependence and spatial heterogeneity.

In recent years, a very extensive literature has stressed the role of spatial effects in many fields of statistical and econometric analysis (Anselin 1988; LeSage and Pace 2009; Benedetti et al. 2015; Kelejian and Piras 2017).

Spatial dependence may be defined as “the propensity for nearby locations to influence each other and to possess similar attributes” (Goodchild 1992, p.33). Empirical models that do not take spatial dependence and structural heterogeneities into account may show serious misspecification problems (see Chapters 9, 12, and 13 in this book).

Spatial dependence may also be referred to as the relationship among outcomes of a variable that is a result of the geographical position of their locations. It measures the similarity of variables within an area and the level of interdependence between the variables (Cliff and Ord 1981; Cressie 1993; Haining 2003). The procedures used to analyse patterns of spatial dependence vary according to the type of data.

For regional scientists, the economic counterpart of spatial dependence is the analysis of spillovers. Measuring the degree of spatial spillovers (LeSage and Pace 2009) and evaluating the extent of contagion (Debarsy et al. 2017) might help policy makers reach a more accurate comprehension of the agricultural phenomena.

The spatial analysis is often based on the definition of contiguity links that enable practitioners to entail geographical structures. These are defined in terms of a proximity matrix (the so-called $W$). Typically, the weights from the $W$ matrix are non-stochastic and exogenous. The weights matrix is often defined with two areas defined as neighbours if they share a common border. In some cases, there is a need to capture and to model other forms of spatial proximity as hierarchical dependence and patterns of spatial competition (Haining 1990). However, another possible approach to define the elements of the weight matrix is in terms of similarity of one or more covariates (Conley and Topa 2002).

Surprisingly, in the spatial econometric and statistic literature, spatial heterogeneity, which is another relevant characteristic highlighted by spatial data, has been less investigated.
Spatial heterogeneity is connected to the absence of stability and it implies parameters to vary over space (Anselin 1988). The presence of a not constant relationship between a response variable and the covariates on a spatial unit has led to the introduction of spatially varying coefficients (Wheeler and Calder 2007), the geographically weighted regression (Fotheringham et al. 2002), the Bayesian regression models with spatially varying coefficient (Gelfand et al. 2003), and the local linear regression models (Loader 1999). In the field of linear estimation, spatial heterogeneity could lead to a serious problem of misspecification of the model (Postiglione et al. 2013).

The spatial heterogeneity can be classified into discrete heterogeneity and continuous heterogeneity. Continuous heterogeneity specifies how the regression coefficients change over space, as estimated, for example, through a local estimation process, as in the geographically weighted regression (GWR, Fotheringham et al. 2002). Discrete heterogeneity consists of a pre-specified set of spatial regimes or a predetermined group of spatial units (Anselin 1990; Postiglione et al. 2013), between which model coefficients are permitted to vary. For further details on spatial heterogeneity applications on agricultural data, see Chapters 8 and 9 in this book.

As highlighted by Postiglione et al. (2017), a new direction in the field of spatial analysis will be represented by the joint treatment of the two spatial effects: spatial dependence and spatial heterogeneity.

1.2.2 Types of spatial data

Spatial data refers to an observation on which we know the value of the variable and the location.

The variables, for example, may be univariate or multivariate, categorical or continuous. They may be based on an observational study, a well-designed experiment, or a sample survey.

The spatial domain, defined as the set of geographical coordinates, offers a potentially huge source of information for the process analysis. There are many different types of spatial data, and, as a consequence, different forms of spatial statistics are required.

According to the classification suggested by Cressie (1993), three types of spatial data can be identified:

- continuous or geostatistical data;
- lattice or areal data;
- point data.

In this classification, spatial data are distinguished by the nature of the spatial domain and not the size of the geographical unit under investigation (Cressie 1993; Schabenberger and Gotway 2005). The fundamental difference is so in the process that generated the data.

First, consider a spatial process in $d$ dimensions as

$$\{x(z): z \in D \subset \mathbb{R}^d\}$$  \hspace{1cm} (1.1)
where \( x \) represents the agricultural variable under investigation, detected at a location \( z \), defined using a \( d \times 1 \) vector of coordinates. If \( d = 2 \) the usual spatial process in a two-dimensional space is defined.

The geostatistical data is defined using a continuous domain \( D \). In essence, with continuous data, there is an outcome for the variable of interest at any point across the territory under investigation \( D \). The continuity is a property of the domain \( D \), not of the variable being measured.

However, these data are usually measured in a discrete number of points defined in a bidimensional space in terms of the Cartesian coordinate \( z_i = (r_i, s_i) \). The point sampling may be chosen according to some design (e.g., random, stratified random, systematic, and spatial designs). See Chapters 2 and 3 for further details.

The points in \( D \) are non-stochastic. A spatial domain is said to be non-stochastic if it does not change from one realization of the spatial process to the next.

The term geostatistics refers to the analysis of continuous spatial variations. Geostatistics was defined by Matheron (1963) as “the application of probabilistic methods to regionalized variables”. Diggle and Ribeiro (2007) identified three different scientific aims of geostatistics: model estimation (inference about the model parameters), prediction (inference about the unobserved values of the target variable), and hypothesis testing. Geostatistics can be also a valid support for optimising the data sampling plan. For further details about a possible application of geostatistics in yield prediction in agriculture, see Chapter 4.

In lattice data, the spatial domain \( D \) is fixed and discrete. The number of locations may be infinite, but they must be countable. Lattice data is often defined in terms of areal regions. The areas may be irregular in shape or form a regular grid. Remote sensed images are a typical example of spatial regular lattice data. In these images, the area is divided into a series of small rectangles, denoted as pixels. See Chapter 5 for some analysis on remote sensed data in agriculture.

Examples of irregular lattice data are most often based upon a partition of the territory into contiguous zones. In this case, the variables observed on spatial units identified by ZIP code, census tracts, provinces, or administrative regions.

The main investigation for areas concerns the relationship between values of neighbouring observations. The spatial analysis of areal data begins with the definition of the neighbourhood structure of the observations, then continues to measure the influence that observations have on their neighbours, and finally evaluates the significance of this influence. Examples of agricultural analyses on areal data are contained on Chapters 11 and 12 of this book.

The domain of geostatistical or lattice data is non-stochastic. Conversely, in spatial point patterns, the set of points changes according to each realization of the random process.

We are faced with point pattern data when each value of the variable refers to the location of a discrete object, the size of which is sufficiently small relative to the study area that it can be treated as a point (e.g., the location of agricultural farms or trees in a region).
The points may have additional information, called marks. In this case, we define the pattern as marked spatial point pattern; otherwise, it is denoted as unmarked.

In essence, spatial agricultural analysis may focus on how the points are distributed within the region (e.g., are the farms spatially clustered or random located?). In this case, the aim is to quantify the gap between the spatial distribution of observations and a completely random distribution in space. If the data presents a pattern more aggregated than a casual configuration, groups of points can be identified, and their importance measured.

Conversely, if we have a marked point pattern, the study may concern spatial analysis of the geographical location and the attribute available (e.g., how is the revenue of clustered/regularly located farms?). Point pattern analyses on agricultural data are presented in Chapter 7.

1.3 Using R software

The R software (http://www.r-project.org) is a statistical environment for the manipulation, analysis, and graphical representation of data. It is an interactive environment, wherein the commands produce an immediate response and provide for object-oriented programming.

The R software provides:

- an efficient data handling and storage capability;
- a collection of operators for calculations on arrays and matrices;
- a large and coherent collection of tools for statistical data analysis;
- graphical facilities for data analysis and display;
- a well-developed, simple and efficient programming language.

R was developed by Ross Ihaka and Robert Gentleman from the University of Auckland in the mid ‘90s’ on the basis of the programming language S and another better-known commercial environment, S-PLUS (Ihaka and Gentleman 1996). The first official stable version was 1.0, released in 2000.

Unlike S-PLUS, this is a GNU-Software, that is, freely available under the constraints of the GPL (General Public License). R is available for Unix/Linux, Windows, and Macintosh platforms, and constitutes a real programming environment.

The codes are distributed through mirror sites, after compliance with the project standards has been checked in relation to the documentation as well. The analyses in R are organized in libraries (packages) which are implemented directly by the software developers. Packages are a collection of functions that are related to each other. Once R is installed, basic packages which only allow standard statistical analysis are immediately available, the other additional packages can be downloaded from mirror sites.

When you use the R program, it issues a prompt when it expects input commands. The default prompt is ‘>’.

The use of R has been simplified by the presence of R-studio (https://rstudio.com/products/rstudio/). RStudio is an integrated development environment (IDE) for R. It incorporates a console, syntax-highlighting editor, tools for plotting, history,
debugging, and workspace management. RStudio is available in open source and commercial editions for Windows, Mac, and Linux, or in a browser connected to RStudio Server or RStudio Server Pro for Debian/Ubuntu, Red Hat/CentOS, and SUSE Linux.

The main problem for the R new users, however, is the knowledge and understanding of statistics. Differently from other commercial statistical software, where the lists of statistical methods are showed through windows or drop-down menus, R requires a priori understanding of the method that should be used. At a first glance, this may seem a huge drawback for using this software; but, in our opinion, this awareness moderates the possibility of a not appropriate use of statistical modelling.

In particular, R is a powerful computing tool that also supports geographical analysis and mapping for researchers interested in spatial analysis and mapping.

This book provides an introduction to the use of R for spatial statistical and econometric analysis for agricultural data. As aforementioned, the choice of using R as the statistical environment for these analyses derives from its particular properties.

1.4 The description of the book

One of the aims of this book is to stimulate and suggest research on the application of spatial statistical methods on agricultural data. The plan of the volume has been conceived to this end.

The book contains, in addition to this introductory Chapter, 12 chapters that attempt to cover multiple aspects of spatial analysis in agriculture. We are aware that this book cannot be considered an exhaustive and definitive reference for the topic, but we believe that the intentionally applicative approach and the extensive use of the R software may be very useful for a large audience of practitioners.

The first phase of a statistical analysis consists of data collection. Chapters 2 and 3 are devoted to this important issue.

In particular, Chapter 2 contains a review of spatial sampling designs. As noted, a spatial population, e.g., those represented by agricultural data, is characterized by dependence between units. As a consequence, neighbour units are likely to provide similar information and this aspect should be carefully considered in the definition of a sampling design. Since standard sampling designs do not consider the influence of spatial proximity between units, several spatial sampling designs have been introduced in the literature and described in this chapter. The R packages used are sampling, Spbsampling, spsurvey, and BalancedSampling.

Chapter 3 presents some issues related to the estimation phase of a survey. The main assumption is that data is coming from complex sampling design. The objective is to make design-based inference for a set of descriptive parameters of a finite population. Also, superpopulation models are considered, since the interest is in accounting for the spatial structure that very likely characterizes agricultural and environmental populations. Thus, a model-assisted approach is used to make inference (Särndal et al. 1992). The description of the estimation phase of sampling designs is supplemented by an illustration on real data coming from a survey on lakes in the Northeastern states of the US, using the R packages sampling and survey.
Chapter 4 describes the topic of yield prediction in agriculture, providing a comparison between the methods of regression kriging and random forest. Crop yields and prediction are becoming crucial activities for farmers as well as for consultants and agricultural organizations. For example, yield forecasting represents an important tool for developing agriculture operations and management and for regulating agriculture cultivation systems. The empirical analysis describes the application of regression kriging and random forest in the spatial estimation of winter wheat yield in the Southern Great Plains of the US, spanning across three states: Kansas, Oklahoma, and Texas. The \texttt{R} packages \texttt{caret} and \texttt{gstat} are used in the illustration.

Chapter 5 contains an interesting description of the problem of land pattern recognition. The production of land cover databases is noteworthy in the last few decades. These tools have been produced using aerial photos or satellite data. They are very important for managing agriculture food security, natural resources, and environment protection. The chapter also analyses different types of methods for the classification of remote sensed data. The application presents a comparison between different classification methods on Sentinel satellite data from Copernicus project of the European Space Agency and ground data provided by the Italian Ministry of Agriculture, using the \texttt{R} packages \texttt{randomForest}, \texttt{adabag}, and \texttt{glmnet}.

Chapter 6 describes different statistical systems in agriculture. The crucial information on agriculture is provided by the Census of Agriculture that is a key pillar for all national agricultural statistical systems. In some cases, this is the only information on the agricultural sector produced by many developing countries. The role of FAO and Eurostat in producing agricultural data is also emphasised in the chapter. This part of the book represents an interesting guide to different available agricultural data.

Chapter 7 presents the introductory aspects of point pattern analysis applied to agricultural data. In agricultural studies, the points, for example, could represent trees, farms, and animal nests. Using spatial point pattern analysis in agriculture, we may study, for instance, the distribution of farm sizes, and the reason why land concentration occurs in a given area. To this end, this chapter describes the introductive aspects of explorative analysis of spatial point patterns and offers a comprehensive description of possible applications on agricultural farms. The empirical exercise is performed using the \texttt{R}-package \texttt{spatstat}.

Chapter 8 focuses on detecting and modelling spatial effects in the analysis of the agricultural production at fixed locations, corresponding to farms. The agricultural output is modelled through a standard Cobb-Douglas production function. Exploratory spatial data analysis tools are presented in order to identify specific patterns of the points. Also, the technique of geographically weighted regression is described as a way to model continuous spatial heterogeneity and to develop local models. The main \texttt{R}-packages used in the application are \texttt{spdep} and \texttt{GWmodel}.

Chapter 9 analyses some aspects related to the spatial econometric modelling of farm data. The definition of the production function and, particularly, the spatial stochastic frontier model are introduced and estimated, using the \texttt{R}-packages \texttt{Benchmarking} and \texttt{ssfa}. Besides, a possible method to identify spatial regimes and contiguous spatial clusters are described for controlling unobserved spatial
discrete heterogeneity. The library spdep and the function skater are used to this end.

Chapter 10 contains an overview of the most common areal interpolation methods. Areal interpolation aims at estimating variables for a set of target zones, based upon the known values in the set of source zones. The main focus is on the Bayesian interpolation method (BIM) introduced by Benedetti and Palma (1994). This technique will be illustrated as a special case of areal interpolation methods that accounts for spatial autocorrelation. The BIM is used to disaggregate data on the production per hectare of firms specialized in olive growing in Italy. The information is available at the regional level and estimated at the provincial level. The R-code to perform the procedure is available in the text.

Chapter 11 is devoted to a description of the topic of small area estimation. This includes the methods for obtaining more precise estimators in local/small areas, making use of the common characteristics of the areas. In particular, the authors analyse the most important area level models, which consider the spatial information, to provide estimates of agricultural and rural statistics at a local/small area level. The theory is complemented by R-illustration using a data set available in the sae R-package: the grape data set. This data set is based on the Italian Agricultural Census of year 2000 for the Italian region of Tuscany.

Chapter 12 presents research focusing on agricultural convergence. The β-convergence refers to a dynamic process in which poorer regions catch up with the richer ones. The importance of this analysis is justified by the relevance of agricultural policies. A cross-section application to β-convergence models for the agricultural sector in Europe is provided. The analysis focuses particularly on the consequences of spatial effects and their treatment in the R environment. To this end, several spatially augmented models are estimated using spdep and spatialreg R-packages.

Finally, Chapter 13 aims at contributing to the literature on spatial panel models, evidencing the main potential of analyses when applied to agricultural data. In particular, the chapter discusses the modern theoretical contributions on spatial panel model, with an illustration of the panel approach to stochastic frontier analysis for agricultural data. The R codes for the estimation are presented and commented on using the library splm. The empirical application concerns the analysis of production efficiency in Indonesian rice farming through the RiceFarms dataset, available in splm.

References


CHAPTER 2
Spatial Sampling Designs
Francesco Pantalone¹,* and Roberto Benedetti²

2.1 Introduction

Usually, a spatial population (i.e., units distributed over a region of interest) is characterized by dependence between units. This is due to the fact that those units are influenced by the same set of factors, which in turn could affect our hypothetical variable of interest. This is especially true in agricultural surveys, where units close together are influenced by the same soil fertility, weather, pollution, and other spatial factors. Correspondingly, contiguous units are likely to be similar and this aspect should be carefully considered in the definition of a sampling design.

A sampling design is a method to select a portion of the population of interest. Technically, given a finite population of interest \( U = \{1, \ldots, N\} \), a sampling design is a probability function \( p(s) \) that assigns a selection probability to each possible subset of the population \( U \), which we call sample and we indicate by \( s \). The set of samples that have a positive probability to be selected by a sampling design is called support of the sampling design. The probability function \( p(s) \), in turn, defines the probabilities of the units in the population to be included in the sample. In particular, the probability that unit \( i \) is included in the sample is called first-order inclusion probability, denoted by \( \pi_i = \sum_{s \ni i} p(s) \), and the probability that units \( i \) and \( j \) are included jointly in the sample is called second-order inclusion probability, denoted by \( \pi_{ij} = \sum_{s \supset \{i, j\}} p(s) \). For more details on the principles of sampling designs see Tillé and Whilelm (2017). It is important to underline the difference between the sampling design \( p(s) \), and the sampling algorithm. With the latter, we implement the sampling design and select samples that respect the probabilities given by the \( p(s) \). In doing so, the algorithm usually uses selection probabilities during some steps, which are not to be confused with the already treated inclusion probabilities. For a deep treatment of sampling algorithms see Tillé (2006).

¹ University of Perugia, Department of Economics.
² “G. d’Annunzio” University of Chieti-Pescara, Department of Economic Studies.
Email: benedett@unich.it
* Corresponding author: francesco.pantalone@studenti.unipg.it
Since standard sampling designs do not consider the spatial dependence between the units, several spatial sampling designs have been introduced in the literature over the last few years. For a review, see Benedetti et al. (2015, 2017c).

In this chapter, we focus on spatially balanced sampling designs, which are sampling designs that achieve samples well spread over the population of interest. The theoretical framework we consider is the design-based approach, which assumes that the variable of interest $y$ is fixed and the only source of randomness comes from the selection of the sample. Moreover, we suppose we are interested in the total of a quantity of interest, say $t = \sum_{i=1}^{N} y_i$. In this framework, a well-known estimator for the total is the Horvitz-Thompson estimator (Horvitz and Thompson 1952), defined as

$$\hat{t}_{HT} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} \quad \text{with} \quad \pi_i > 0 \ \forall \ i \in U.$$ 

Its variance is equal to:

$$V(\hat{t}_{HT}) = \sum_{U} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j,$$

and is estimated by

$$\hat{V}(\hat{t}_{HT}) = \sum_{U} \sum_{ij} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j.$$

The use of spatially balanced designs could improve the efficiency of the final estimates, in the sense that the obtained variance of the HT estimator could be lower with respect to the variance obtained by the use of some non-spatial sampling designs.

The outline of the chapter is as follows: the first sections are dedicated to some practical aspects, specifically the frame in Section 2.2 and the process of data collection in Section 2.3. Both of these aspects are important in the survey process, because they could lead to problems if not treated correctly. In Section 2.4, we provide a brief introduction to non-spatial sampling designs, specifically the simple random sampling, widely used as benchmark and as part of more sophisticated sampling designs; the systematic sampling, especially useful for its simplicity; the stratified sampling, which is a simple and efficient way to consider some structure of the population at the design level, and the CUBE method, for balanced samples. In Section 2.5, we focus on spatially balanced sampling designs. In particular, we analyse the generalized random tessellation stratified sampling (Steven and Olsen 2004), based on a function that maps two-dimensional space onto a one-dimensional space; the spatially correlated Poisson sampling (Grafström 2012), adaptation of Correlated Poisson Sampling (Bondesson and Thorburn 2008); the local pivotal method (Grafström et al. 2012), modification of pivotal method (Deville and Tillé 1998); product within distance and sum within distance (Benedetti and Piersimoni 2017a), based on a MCMC algorithm and a summary index of a distance matrix. Finally, we provide some concluding remarks.

### 2.2 Setting up the frame

An important task of any survey is the identification of the target population (e.g., people aged over 40, firms with more than 50 employees, etc.). Once the
target population is identified, the next step is to gather information about units. In particular, a list of units is needed in order to perform a probability sampling design. We collect such information through the sampling frame, which is defined as any material or device used to obtain observational access to the finite population of interest (Särndal 1992). In the following, we list some of the properties a sampling frame should have (Colledge 2004):

- identification of the units through the use of a unique identifier code;
- completeness, i.e., the frame contains all the units in the population;
- accuracy, i.e., there are no repetitions of units;
- additional information about the units (when possible).

The quality of the sampling frame has a large impact on the quality of any survey. In agricultural survey, defective sampling frames are a common source of non-sampling error. Therefore, definition and data collection of the frame is an important task and should be carried out with particular attention. To this end, we provide some guidelines for the definition and the construction of the frame. Roughly speaking, we can define two main steps: (i) definition of the statistical unit; and (ii) definition and construction of the frame. In the following, we analyse these steps with particular focus on agricultural surveys.

(i) Definition of the statistical unit

We can broadly divide statistical units into two categories:

- **legal bodies**, such as farms, households and businesses. Agricultural surveys are usually based on farms, but households and businesses are frequently used as well;
- **spatial units**, which are portions of land areas and they could be points, polygons, and lines. In agricultural surveys, points and polygons are of the main interest. These portions of areas can be set by identifiable physical boundaries, such as rivers or roads, using squared grid of map coordinates or superimposing their limits to the boundaries of the land of the agricultural holding.

(ii) Definition and construction of the frame

According to the type of statistical unit, a different frame is employed. In particular, when legal bodies are used as statistical units, list frames are defined, while in the case of spatial units, spatial frames are considered. The former is a list of agricultural holding addresses, along with other information, such as holding size, crops, livestock, and possibly more. The addresses can be obtained from previous agricultural censuses, administrative data sources, and farmer associations (Wallgren and Wallgren 2007; 2010). Among administrative data sources, it is worth mentioning the so-called cadastre, which is the up-to-date land records, usually owned by country institutions. Moreover, it is important to consider holdings under different legal status, such as cooperatives, government farms, and enterprises (FAO 1995).

A spatial frame is a list of non-overlapping areas. The preparation of this type of frame is particularly demanding and requires up-to-date graphical materials, such as maps, satellite images, and aerial photos.
2.3 Survey data processing

Once the units have been selected, the process of data collection begins. More specifically, we can identify three major processes: data collection, data editing, and validation process. In the data collection process, the selected units are observed. The main objectives of this process are three: (i) identification of statistical unit (e.g., farm, polygon, point); (ii) collection of data without introduction of endogenous bias (e.g., avoid to influence the response); (iii) consideration of all stages of the sampling and eventually a longitudinal structure in order to facilitate any future contact. Once the data is collected, it is necessary to detect any eventual missing or inconsistent data and account for it through data imputation. This is the task of the data editing process, which is usually fully automatic or computer-assisted (Atkinson and House 2010). Finally, a validation process evaluates the consistency and the quality of the data, and it is composed by all operations required to check if the achieved results meet the planned quality targets. Not only, this validation assesses if the quality of the data is sufficient for distributing the data to users, but it could also identify possible sources of error. Therefore, this process could determine relevant modifications to the future surveys in order to reduce eventual future errors.

2.4 Non-spatial sampling designs

In this section, we briefly review some of the non-spatial sampling designs. With the term non-spatial, we mean that these sampling designs have not been designed specifically for the use on spatial settings. Nevertheless, they can be extended to that use in some situations, and we provide insights on how to apply them, if possible. We first consider the Simple Random Sampling, the Systematic Sampling, and the Stratified Sampling. Such designs are widely used in practice, as standalone or as part of more complex sampling designs. Then we introduce the CUBE method, which is a way to select balanced samples. For each case, we provide R code in order to perform sampling selection. In doing that, we use a simulated population $U$ available in the dataset simul2 from package Spbsampling (Pantalone et al. 2020).

2.4.1 Simple random sampling

From a randomization point of view, simple random sampling (SRS) is one of the most basic sampling designs. In fact, given $n$ as the sample size and $N$ the population size, the SRS assigns to each sample of dimension $n$ the same probability to be selected, given by $p(s) = \frac{1}{\binom{N}{n}}$.

When speaking of SRS, we need to differentiate between two types: with replacement (SRSWR), and without replacement (SRSWOR). In the former, the unit selected in a draw is then re-inserted in the population, allowing the possibility to have the same unit more than once in the sample; in the latter, once the unit is selected, it cannot be re-selected. The most widespread method in use is SRSWOR, both for reasons of efficiency, since the expected number of distinct units in the sample is constant and equal to $n$, and therefore variance of estimators of population total (and mean) is smaller compared to when using SRSWR; and for practical
reasons, because in many surveys we may not want to observe the same unit more than once. Therefore, we focus on SRSWOR and we refer to it simply by SRS.

In this scenario, every unit in the population has the same first-order inclusion probability, constant and equal to the sampling fraction $f$,

$$\pi_i = \frac{n}{N} = f \forall i \in U,$$

while the second-order inclusion probabilities are equal to

$$\pi_{ij} = \frac{n}{N} \cdot \frac{n-1}{N-1} \forall i \neq j \in U.$$

Thus, the HT estimator of the total is given by

$$\hat{y}_{HT,SRS} = \sum_{i=1}^{N} \frac{y_i}{\pi_i},$$

and the variance estimate by

$$\hat{V}_{HT} (\hat{y}_{HT,SRS}) = N^2 \left( \frac{1}{n} S_y^2 = N^2 \left( \frac{1}{n} \frac{1}{N} \right) S_y^2 \right),$$

where $S_y^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$.

By definition, SRS does not allow the use of auxiliary information in the design phase, therefore, no kind of population structure can be exploited. Indeed, the use of SRS is recommended when, for instance, the population lacks hierarchical or geographical structure. Also, many more complex sampling designs use in some steps a simple random sampling.

From a methodological point of view, SRS provides important insights on the analysis of complex sampling designs. In fact, it usually represents a benchmark in design effects comparison. Given $\hat{V} (\hat{y}_{HT,p})$ as the estimated variance of the HT estimator of the total when a sample of size $n$ obtained by a generic sampling design $p$ is used, and $\hat{V} (\hat{y}_{HT,SRS})$ is the estimated variance of the HT estimator of the total when a SRS of size $n$ is used, the design effect is defined as the following variance ratio

$$deff (\hat{y}_{HT,p}, \hat{y}_{HT,SRS}) = \frac{\hat{V} (\hat{y}_{HT,p})}{\hat{V} (\hat{y}_{HT,SRS})}.$$

When this ratio is lower than 1, the use of the sampling design $p$ leads to a gain in efficiency with respect to the use of SRS, otherwise there is a loss of efficiency. In addition, the design effect can be used to analyse different couples of estimators and sampling designs (usually referred as sampling strategy). Furthermore, following this approach, we could compare the efficiency of different estimators, given the use of the same sampling design.

We can perform SRS in the R environment by means of the function `srswor()` of the package `sampling` (Tillé and Matei 2016). The function requires as inputs the parameters $n$ and $N$, which are the sample size and the population size, respectively.
The output is a vector of size $N$, where the $i$-th element is equal to 1 if the $i$-th unit is selected in the sample, 0 otherwise. As an example, we select a sample of dimension 100 through the following code.

```r
# -- selection of the sample -- #
>N <- nrow(simul2) # population size
>n <- 100 # sample size
>pi <- rep(n/N, N) # inclusion probabilities
>set.seed(42)
>srs_sample <- which(srswor(n, N) == 1)
```

### 2.4.2 Systematic sampling

Systematic sampling consists of selecting sample units from a list of population units using a constant skip, i.e., every $r$-th unit is selected in the sample, where the start is random in order to achieve random samples. The list plays a crucial role. If it is shuffled randomly before the selection, the systematic sampling can be compared with SRS from a theoretical point of view. If it is ordered according to some criteria, the efficiency of the selection is greatly influenced by that. For example, if a population of agricultural firms is ordered according to yield crop, then the sample would reflect the yield crop distribution of the population better than by chance. With that being said, the use of systematic sampling can often encounter the practical disadvantage of a ratio $N/n$ not integer, which implies that it is often impossible to find a step $r$ such that we sample exactly $n$ units (in univariate populations, a solution is proposed in Särndal et al. (1992, p. 6)). Moreover, due to the sampling mechanism, the majority of second order inclusion probabilities are equal to zero. Therefore, the standard estimator for the variance of HT cannot be used. For some possible solutions, see Särndal et al. (1992, p.83). The most used workaround is to assume as upper bound of the variance an SRS variance estimate in order to provide a cautionary estimate.

Systematic sampling can be used to consider spatial aspects of the population. In fact, using the coordinates of the units in the ordering phase could lead to a good spatial coverage. Moreover, to improve and increase the use of the spatial information, we can divide the population into equally spaced clusters. Cochran (1997) showed that this design has a lower variance than SRS if $\rho_{IC} < -n/[N(n - 1)]$, where $\rho_{IC}$ is the correlation between a pair of units in the same systematic sample, or intraclass correlation coefficient. The main difficulty when using the systematic design in a spatial context is given by the fact that only a regular grid of points or a set of regularly shaped polygons have a natural ordering that can be used to spread the sample over the study region.

The following code selects one sample of dimension 100.

```r
# -- selection of the sample -- #
>N <- nrow(simul2) # population size
>n <- 100 # sample size
>pi <- rep(n/N, N) # inclusion probabilities
>set.seed(42)
>sys_sample <- which(UPsystematic(pi) == 1)
```
2.4.3 Stratified sampling

Taking into account the population structure during the selection process is an important task that can greatly benefit the efficiency of the final estimates. This process requires good auxiliary information, and in stratified sampling such information is used in order to create subgroups of the population, which are called strata. Afterwards, a sample is independently selected from each stratum. The auxiliary information should be used in such a way that the created strata have differences between them but similar on their inside. Indeed, it has been shown that stratified sampling is more efficient than SRS when the units within each stratum are as similar as possible, and the units in different strata are as different as possible (Cochran 1977).

The population $U$ is partitioned into $H$ strata, say $\{U_1, U_2, \ldots, U_h, \ldots, U_H\}$, which are exhaustive, i.e., $\bigcup_{h=1}^{H} U_h = U$, and non-overlapping, i.e., $U_h \cap U_r = \emptyset \forall h \neq r$. Therefore, given $\{N_1, N_2, \ldots, N_H\}$ be the numbers of units of the population belonging to each stratum, then $\sum_{h=1}^{H} N_h = N$. From each stratum, a sample of dimension $n_h$ is selected. Theoretically, any sampling design can be employed when sampling from the strata, and even the use of different sampling designs across the strata is allowed. Regarding point and variance estimation, since the random selection inside a stratum is independent from every other random selection in the other strata, the HT estimator and corresponding variance estimator are given simply by the sum of the estimates in each stratum. This allows a great deal of flexibility, because different sampling designs can be employed across the strata without the need of face complications on the estimation phase.

Here, we consider that an SRS is selected from each stratum. Therefore, first-order inclusion probability for unit $i$ in stratum $h$ is equal to

$$\pi_i = \frac{n_h}{N_h} = f_h \forall i \in U_h$$

and second-order inclusion probabilities are given by

$$\pi_{ij} = \frac{n_h}{N_h} \frac{n_h-1}{N_h-1} \forall i \neq j \in U_h \text{ and } \pi_{ij} = \frac{n_h n_m}{N_h N_m} \forall i \in U_h \forall j \in U_m, h \neq m.$$ 

Hence, we have the HT of a total as in the follow

$$\hat{t}_{HT,STR} = \sum_{i=1}^{H} \frac{y_i}{\pi_i} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in h} y_i$$

and the variance estimator equal to

$$\hat{V}_{HT} (\hat{t}_{HT,STR}) = \sum_{h=1}^{H} N_h 2 \frac{1-f_h}{n_h} S_{y,h}^2,$$

where $S_{y,h}^2 = \sum_{i \in h} (y_i - \bar{y}_h)^2$.

Another choice to face with stratified sampling is size allocation, i.e., how to determine $n_h$ for each stratum. Two approaches are Neyman Allocation and Optimal Allocation. Here, we just give the idea behind these two methods and we refer to
Särndal (1992) for a deep treatment of the subject. Neyman allocation minimizes the variance of a sample estimate subject to a given sample size, and under this allocation \( n_h \) should be proportional to \( N_h S_h \), where \( S_h \) is the population standard deviation of stratum \( h \). Optimal allocation minimizes the variance of a sample estimate subject to a given total cost, and \( n_h \) should be proportional to \( N_h S_h / \sqrt{C_h} \), where \( C_h \) is the cost of observing a unit \( i \) in the stratum \( h \).

Spatial auxiliary information can be used in order to stratify the population, even though it is not always clear how to efficiently perform such task. A common solution is to define the strata as subareas of the region of interest.

The package `sampling` (Tillé and Matei 2016) provides the function `strata()`, which performs a stratified sampling according to the main following parameters:

- `data`: population frame;
- `stratanames`: vector of stratification variables;
- `size`: vector of stratum sample sizes;
- `method`: method used inside the strata in order to select the units. The methods implemented are sampling without replacement (`srswor`), simple random sampling with replacement (`srswr`), Poisson sampling (`poisson`), systematic sampling (`systematic`).

In the following, we create two strata from the dataset, then select a stratified sample, and from each stratum we select an SRS.

```r
# -- selection of the sample -- #
>N <- nrow(simul2)  # population size
>n <- 100  # sample size
>pi <- rep(n/N, N)  # inclusion probabilities
>stra <- as.numeric(simul2$z21 > 5)
>f <- cbind(simul2, stra)
>set.seed(42)
>str_sample <- strata(data = f, stratanames = "stra", size = c(50, 50), method = "srswor")
```

### 2.4.4 The CUBE method

Balanced sampling can be viewed as a calibration method employed into the selection process, and it takes advantage of auxiliary information. Given \( \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \) as the vector of \( p \) auxiliary variables measured on unit \( i \), a sample is called balanced with respect to those variables if the HT estimates of the totals are equal to the known totals \( t = \sum_{i \in \mathcal{U}} x_i \), that is

\[
\sum_{i \in \mathcal{E}_i} \frac{x_i}{\pi_i} = t.
\]  

(2.1)

If auxiliary variables used are correlated with the variable of interest, a balanced sample could lead to gain in efficiency estimates. Note that balanced sampling generalises other sampling designs. For example, a sampling design with fixed sample
size $n$ is balanced on the variable $x_i = \pi_i$, while a stratified sampling is balanced on the indicator variables of strata. The CUBE method (Deville and Tillé 2004) selects balanced samples with any number of auxiliary variables, while a set of prescribed first-order inclusion probabilities is satisfied (even unequal probabilities are allowed). The algorithm starts from the vector $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$, and at each step at least one component is changed to 0 or 1. This change is made in such a way that the prescribed inclusion probabilities are satisfied, and the balancing equations (2.1) are satisfied as well. In particular, the method is composed of two phases. In the first one, the flight phase, the balancing equations are satisfied exactly. The second one, the flight phase, takes place in the eventuality that in the flight phase no sample satisfies the equations precisely, and it slightly relaxes the constraints as consequence. In this case we achieve a sample that approximately satisfies the balancing equations (2.1).

The implementation in the R environment is provided by the package BalancedSampling (Grafström and Lisic 2016), through the function cube(). Parameters needed are prob, vector of first order inclusion probabilities, and Xbal, matrix of auxiliary balancing variables. Note that, in order to have a fixed sample size $n$, we need to insert the first order inclusion probabilities that sum to $n$ as first column of the matrix Xbal. The output of the function is a vector containing the labels of the units selected. In the following, we select a sample by the CUBE method using the spatial coordinates of the units as balancing variables.

```r
# -- selection of the sample -- #
>N <- nrow(simul2) # population size
>n <- 100 # sample size
>pi <- rep(n/N, N) # inclusion probabilities
>X <- cbind(pi, simul2$x, simul2$y)
>set.seed(42)
>cube_sample <- cube(prob = pi, Xbal = X)
```

### 2.5 Spatially balanced sampling designs

Over the last decade, spatial data has increased substantially. New technologies allow geo-coded data to be obtained easily and, consequently, official statistical offices have started to geo-reference their registers. When dealing with spatial populations, we should consider the spatial structure. In fact, units that are spatially distributed usually show spatial dependence. In agricultural research, surveys are routinely used to gather data. The observed units are typically geo-referenced and it would, therefore, be desirable to take into account the spatial distribution when selecting the sample. **Spatially balanced sampling designs** address the task of taking into consideration the spatial dependence of the units in the population. In particular, these designs select samples well spread over the population of interest. Technically, a well spread sample has a number of selected units on every part of the study region close to what is expected on average (Grafström and Lundström 2013). The idea is that a spread sample could capture the spatial heterogeneity of the population, which in turn could improve the efficiency of estimates compared to the efficiency of estimates achieved by data obtained from non-spatial sampling design.
(e.g., SRS). We review two theoretical reasons that clarify this idea, that are the *lemma decomposition* and the use of the *anticipated variance*.

The decomposition lemma (Knottnerus 2003, p.87) states that, given a sample of size $n$ drawn by a sampling design $p(s)$, the constant and unknown population variance of the variable $y$ is given by

$$
\sigma_y^2 = V_S(\tilde{y}_S) + \frac{n-1}{n} E_S(S^2_{\tilde{y},S}),
$$

where $\tilde{y}$ is the expanded vector of $y/\pi$, $V_S(\tilde{y}_S)$ is the variance between samples of HT estimator of the mean according to the design $p(s)$ and $E_S(S^2_{\tilde{y},S})$ is the expectation of the sample variances of $\tilde{y}$. From this result, we can observe that HT estimator could be more efficient when the first-order inclusion probabilities are such that the $y/\pi$ are constant or proportional to $y$, and/or the design $p(s)$ increases the expected within sample variance. The latter suggests that the $p(s)$ should be proportional or more than proportional to the sample variance $S^2$. Since this variance is unknown because it is relative to the unobserved variable of interest $y$, we should have auxiliary information related to it in order to exploit this theoretical insight. In the spatial setting, such information can be provided by the distance between units, as pointed out in the spatial interpolation literature (Ripley 1981; Cressie 1993).

Another way to understand the possible consequences of a well spread sample is by modelling the population. To this end, we introduce a spatial model that *(i)* links the variable of interest with some auxiliary information; *(ii)* considers the spatial correlation. From the model, we compute the *anticipated variance* (Isaki and Fuller 1982), from now on AV, of the HT estimator. In particular, the AV of a parameter of interest $\theta$ and its estimator $\hat{\theta}$ is given by

$$
AV(\hat{\theta} - \theta) = E_m\left[E_S\left[(\hat{\theta} - \theta)^2\right]\right] - E_m\left[E_S\left[\hat{\theta} - \theta\right]\right]^2,
$$

where $E_m$ denotes expectation with respect to the model and $E_S$ denotes expectation with respect to the sample design.

The following linear model is assumed to hold for $y$, given the known auxiliary variables $x$, for each unit in the population:

$$
y_i = x_i'\beta + \epsilon_i, \quad E_m(\epsilon_i) = 0,
\begin{align*}
 Var_m(\epsilon_i) &= \sigma^2_i, \\
 Cov_m(\epsilon_i, \epsilon_j) &= \sigma_i\sigma_j\rho_{ij}
\end{align*}
\tag{2.2}
$$

where $Var_m$ and $Cov_m$ denote variance and covariance with respect to the model, respectively, $\beta$ is a vector of regression coefficients, $\epsilon_i$ is a zero-mean random variable with variance $\sigma^2_i$ and $\rho_{ij}$ is an autocorrelation parameter, for $i \neq j$ and such that $\rho_{ij} = 1$ if $i = j$. Under the model (2.2), the AV of the HT estimator of the total $t_y$ is given by (Grafström and Tillè 2013)

$$
AV(t_{y,HT}) = E_S E_m(t_{y,HT} - t_y)^2 = E_S \left[ \sum_{i=1}^n \frac{x_i}{\pi_i} - \sum_{i=1}^n x_i \right] \beta^2 + \sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}, \tag{2.3}
$$
From equation (2.3), we observe that the uncertainty about the estimate can be split into two terms. The first term can be reduced selecting a sample that is balanced on $x$, while the second term of the equation can be reduced exploiting the spatial information of the population. Indeed, if $\rho_{ij}$ increases as the distance between the units decreases, then selecting units far apart reduces the second term. Therefore, a sample well spread over the region of interest could potentially improve the HT estimation.

For the HT variance estimation, we cannot rely to the general variance estimator presented in the introduction since these sampling designs usually set the second order inclusion probabilities of nearby units to zero. Moreover, the computation of such probabilities can be prohibitive. To overcome this problem, we can employ the local mean variance estimator (Steven and Olsen 2003), or a generalization of that estimator by Graaström and Schelin (2014). Indeed, both proposals do not involve the use of the $\pi_{ij}$s. With a different approach, Benedetti et al. (2017b) proposed a model-based estimation of the variance in non-measurable designs.

The spatial balance of a sample can be measured by the spatial balance index (SBI, Steven and Olsen 2004), which is an index based on the use of Voronoi polygons. For a given sample $s$, a Voronoi polygon is constructed for each sample unit $i$, and it includes all population units closer to $i$ than to any other sample unit $j$. For the $i$-th Voronoi polygon, $v_i$ is defined as the sum of the inclusion probabilities of the units it contains. The index is then

$$ SBI(s) = Var(v_i) = \frac{1}{n} \sum_{i=1}^{n} (v_i - 1)^2. $$

Since for a spatially balanced sample $v_i \approx 1 \ \forall i \in s$ (Steven and Olsen 2004), then the closer $SBI$ is to zero, the better the spread.

In the next sections, we will review the main spatially balanced sampling designs. The theory is presented along with illustrations in R code. For the sake of the illustration, we select samples from the same simulated population $U$ we used in the previous section and, in order to investigate the spatial balance, we compute the corresponding SBI by means of the sbi() function, which can be found in the package Spbsampling (Pantalone et al. 2020). This function requires three parameters: dis, distance matrix that contains all the distances of all the pairs of units in the population, pi, vector of first order inclusion probabilities, and s, vector of labels of the units selected in the sample. Here, we compute the SBI for the samples we have already selected in the previous section. In this way, we can compare the results in term of spatial balance between the non-spatial and the spatial sampling designs.

```r
# -- computation of sbi -- #
>d<-as.matrix(dist(cbind(simul2$x, simul2$y)))
>sbi(dis=d, pi=pi, s=srs_sample)
[1] 0.389899
# -- computation of sbi -- #
>d<-as.matrix(dist(cbind(simul2$x, simul2$y)))
```
>sbi(dis = d, pi = pi, s = sys_sample)
[1] 0.3810101

#-- computation of sbi --#
>d <- as.matrix(dist(cbind(simul2$x, simul2$y)))
>sbi(dis = d, pi = pi, s = str_sample$ID_unit)
[1] 0.4539394

#-- computation of sbi --#
>d <- as.matrix(dist(cbind(simul2$x, simul2$y)))
>sbi(dis = d, pi = pi, s = cube_sample)
[1] 0.4791919

2.5.1 Generalized random tessellation stratified sampling

Generalized random tessellation stratified (GRTS) sampling is the first spatially sampling design introduced in the literature (Stevens and Olsen 2004). Some preliminary works can be found in Stevens (1997), who derived inclusion and joint inclusion functions for several grid-based precursor designs to GRTS, and introduced the multiple-density, nested, random-tessellation stratified (MD-NRTS) design.

Core of GRTS is a function that maps two-dimensional space into one-dimensional space while preserving some spatial relationships. Specifically, the units located within a geographic region are placed on a line by means of a hierarchical randomization process, from where a systematic sample is then selected. In particular, a quadrant-recursive function is used through the ordering process of the units on the line. This process starts with a $2 \times 2$ square grid put randomly over the region of interest, with the resulting cells placed randomly in a line. Then, for each cell the same process is repeated, and the resulting sub-cells are randomly ordered inside the original cell (at this second step, 16 cells are put in a line). The process continues until at most one population unit occurs in a cell, and the generated random order is used to place the units on the line, which therefore has length $N$. A systematic sample of dimension $n$ is then selected from the line. Indeed, the line is divided into $N/n$ length segments, and a starting point $r$ is selected randomly from $(0, N/n)$. Then, every $(r + iN/n)$th point, for $i = 1, \ldots, n-1$, is selected, where if the point occurs within one of the units, then that unit is selected (Brewer and Hanif 1983). If unequal probabilities are prescribed, then each point is given a length proportional to its inclusion probability. This procedure allows to achieve spatially balanced samples that respect given inclusion probabilities. Even though we have focused on area sampling, the GRTS can be implemented for point, linear features and not contiguous phenomena as well. However, there is the possibility to lose some spatial relationships during the hierarchization process, for instance it can happen that units close in the space may be far apart in the one-dimensional space. Finally, note that this method cannot be applied when units have more than two coordinates, i.e., when the space domain is given by $D \subset \mathbb{R}^h$, with $h > 2$.

GRTS is implemented on the R package spsurvey (Kincaid and Olsen 2016) through the function grts(). In order to ease the illustration, we present a function which wraps the original. Inputs are given by $p$, vector of first-order