Politecnico di Milano

DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING

DOCTORAL PROGRAMME IN ENVIRONMENTAL AND INFRASTRUCTURE ENGINEERING

EXPLORATORY APPROACHES IN SPATIAL ASSOCIATION ANALYSIS: METHODS, COMPLEMENTS, AND OPEN GIS TOOLS DEVELOPMENT

Doctoral Dissertation of:
Daniele Oxoli
Matr. 861153

Supervisor: **Prof. Maria Antonia Brovelli**
Tutor: **Prof. Giovanna Venuti**
Chair of the Doctoral Programme: **Prof. Alberto Guadagnini**

XXXI Cycle (2015-2018)

This page intentionally left blank

Acknowledgements

I want to express to:

- My supervisor, Professor Maria Antonia Brovelli for her wise suggestions both in and outside the academia. These have been part of the person I am today.

- My closest colleagues, Marco, Eylül, Monia, Carolina, Gabriele, Giulia, Stefano, Mayra (not necessarily in this order), and all the fellows that shared with me the best and the worst times during these years at PoliMI.

- Professor Abbas Rajabifard, and his team at the University of Melbourne. They welcomed me as a family during my short stay making priceless the time I spent there (also for completing this thesis work).

- My friends, and all the people that I both met and missed during these years. They have always reminded me that life is something more than what I could figure out on my own.

- My family, no word can properly depict what I would like to express to them.

'The greatest value of a picture is when it forces us to notice what we never expected to see.'

John W. Tukey

Abstract

The analysis of complex natural and social systems - using modern geospatial data - requires dedicated methods and tools to grasp their characterising features while accounting for the geographical context where they take place. The same applies both to data discovery and representation. In views of the above, the use of Exploratory Spatial Data Analysis is here leveraged alongside its application into Geographic Information Systems to uncover underlying characters of geospatial data. Among these, the spatial association is considered as the critical aspect to be investigated in this work. A comprehensive review of popular statistical methods for measuring and mapping spatial association is presented together with a description of the most cutting-edge software tools to perform them. A Free and Open Source Software tool dedicated to the spatial association mapping is developed, and its use into sample case studies is discussed. These encompass well-established applications of local spatial association statistics that focus mainly on univariate analysis. To account for both the growing complexity and abundance of the modern geospatial data, extensions of these statistical methods are outlined to enable spatial association analyses in a multivariate context. Experimental results from early applications of these extended methods are disclosed and critically discussed. Finally, the lesson learned and the future directions for the work are presented together with more general considerations on the role of the spatial association in both present and future geospatial data analysis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the era of *Big Data*, the main challenge to address for data scientists and analysts is most definitely to turn data into actionable insights. Data has been recognized as a fundamental asset in manifold sectors of both modern science and business [65]. At the same time, it has been recognized that data alone means nothing without a proper purpose and interpretation [107]. During the last three decades, the Information Technology (IT) has brought terrific advances in data management and tooling but still leaving the duty to scientists of converting data into spendable knowledge [67]. The increasing complexity and richness of the modern *Big Data* have enforced the need for tools and methods to facilitate this latter task. These have to be properly designed to cope with data heterogeneity and volume rather than individual observations quality and accuracy. In other words, these tools have to focus on data *characters* such as underlying patterns, links, and redundancy in order to ease data exploration and understanding [69].

With this in mind, the present work aims at introducing methods as well as suggesting novel solutions - embedding the above concepts - for investigating *characters* of a specific type of data, namely geospatial data. For geospatial

data, we intend all the information describing a physical object or a generic event that can be represented by numerical values in a geographic coordinate system [94]. According to the debatable assertion *"80% of all information is geographic"* [62], much of the data in the world is - or can be - georeferenced. The latter can be easily recognised by thinking to the flow of information which is continuously produced by Earth Observations (EO) satellites, mobile devices, physical sensors, public administrations, private enterprises, etc. which is intrinsically connected to a defined location both in time and space. On one hand, the increasing number of formats, sources, and purposes of the collected geospatial data presents challenges in data storing, managing, deploying, security, and quality [78]. On the other hand, access to and interaction with geospatial data are of the utmost importance to explore natural, human, and social systems [79] and provide outstanding opportunities to gain insights into complex phenomena such as climate change [85], disease surveillance [64], disaster response [28], critical infrastructures monitoring [101], transportation [20], and many others. Despite the majority of these applications are confined into the research and public sectors [35], a growing value has been attributed to geospatial data also by the business sector [27]. Among the business-oriented activities involving the use of geospatial data, the Location Intelligence [14] and the Geomarketing [95] are perfect examples of tasks heavily impacting trades and markets. In general terms, besides that most of the geospatial information can be still handled with the available geospatial theory, tools, and methods, there is an emerging need for establishing *spatial enablement* into general data analysis frameworks for advancing leading-edge research as well as good practices within all those domains exploiting geospatial data [99].

In this work, I leverage the use of Exploratory Spatial Data Analysis (ESDA) [13] as a key tool to enable geospatial data *character* investigations. ESDA can be broadly defined as a framework including a number of statistical methods, vi-

sualisation techniques and software tools to identify spatial patterns and trends as well as to accurately discover and account for spatial relations characterising most of the geospatial datasets [5]. The *character* that is mainly considered along this work is the spatial association.

Together with a review of some methods of interest, which primarily focus on the spatial association analysis of geospatial data with a single or double attributes (i.e. univariate or bivariate analysis), an extension of these methods to account for multiple attributes (i.e. multivariate analysis) is here developed and presented. The latter promises to be relevant for tackling the increasing complexity of the information potentially available at any location on Earth. Therefore, the main advantage of the presented research within Big Data analysis is actually entitled on a single edge of the Big Data that is the variety [47]. This intention is embedded in the output of this work which includes the multivariate extensions of traditional spatial association statistics. Other aspects of Big Data are not tackled by the presented work.

In parallel, some of the relevant software tools to perform ESDA are introduced. These are generally provided as modules of software suites dedicated to geospatial data management and analysis, known as Geographic Information Systems (GIS). Particular attention is paid here to the development of ESDA tools for Free and Open Source Software (FOSS) GIS. FOSS includes all those software which are freely distributed with open licenses allowing users to access, modify and redistribute them for any purpose. In contrast to copyrighted, closed-source proprietary alternatives, FOSS enables free access to the source code in order to favour the application reuse and customisation. This is perfectly aligned with the underlying goal of this research work that is to improve both visibility and usage of ESDA tools among the largest community of data scientists and analyst [86].

Both methods and tools are also presented by taking advantage of applications on real data. For this purpose, heterogeneous geospatial data sources are considered. These range from the Volunteered Geographic Information (VGI) [66] to the georeferenced user-generated content [19] passing through official geospatial data from national bureaux of statistics, economics, and environmental protection. Nevertheless, these represent only a byte of the possible geospatial data sources to which ESDA can be applied.

Focusing on the key points mentioned above, the rest of the document is organized as follows. Chapter 2 contains a deeper introduction to ESDA and the spatial association, including their main features and requirements. Chapter 3 outlines the ESDA methods of interest for this work. Chapter 4 describes the available and most popular ESDA software tools. Chapter 5 presents sample applications on real data. A primer on the application of some ESDA methods in a multivariate context is presented in Chapter 6. In Chapter 7, both key conclusions and the future directions of the work are discussed.

Chapter 2

Exploratory Spatial Data Analysis and Spatial Association

In this chapter, the general features of ESDA are described together with the concept and analytical implications of the spatial association. This is anticipated by a discussion on the meaning of exploratory analysis and its contribution to the modern data science.

2.1 Exploratory Spatial Data Analysis

2.1.1 The exploratory approach in data science

In Chapter 1, we defined ESDA as a framework including a number of statistical methods, visualisation techniques and software tools to identify spatial patterns and trends as well as to properly discover and account for spatial relations of a geospatial dataset. The acronym derives from the more famous Exploratory Data Analysis (EDA) [105] which can be described as the critical process of carrying out data investigations to detect patterns, spot anomalies, test hypothesis, and check assumptions by exploiting mainly descriptive statistics and

graphs. ESDA inherits from EDA the key concepts of dynamic interaction with data and focus on its summarizing characteristics rather than individual observations quality. This by coupling visual and statistical methods to bring out underlying data features. Both EDA and ESDA incorporate most of the traditional statistical methods for data analysis although their role is here marginal or - at most - purely functional to the interactive procedure of formulating and checking assumptions as well as design new experiments on the data [9]. In this exactly resides the meaning of exploratory analysis which has to be distinguished from the confirmatory one. In fact, confirmatory analysis aims at attesting conclusions rather than generate new hypotheses based on the data. Nevertheless, these two approaches should be considered complementary and not competitive within any comprehensive analysis framework. Indeed, the exploratory step should encompass the whole analysis cycle being this valuable both during the preliminary checks on data as well as for the results assessment. In general terms, the principles and the ultimate utility of EDA are summarized in the statement *"It is important to understand what you can do before you learn to measure how well you seem to have done it"* formulated by the pioneer of EDA J. K. Tukey [105].

In practice, one of the main distinction between the exploratory and the confirmatory approach is that the first, according to its goals, emphasises the use of data visualisation tools and techniques. Therefore, most of the existing graphical techniques in data analysis have been developed as a consequence of the principles introduced by EDA. These techniques include popular statistical graphics such as histograms, box-plots, scatter-plots, stream-graphs, table-based plots, and many others [108]. Furthermore, not only graphical techniques but also quantitative methods have been developed to cope with the EDA principles. These methods mainly face up to:

- *Dimensionality reduction*, e.g. the Principal Component Analysis (PCA) [70]

- *Clustering*, e.g. the K-Mean Clustering [38]

- *Uncertainty estimation*, e.g. the Bootstrap [40]

- *Outliers detection and hypothesis testing*, e.g. permutation tests [58].

Most of the graphical and quantitative EDA techniques are at the core of the modern computational statistics, which represents the interface between statistics and computer science. The computational statistics is nowadays crucial in numerous scientific fields and increasingly adopted also in the business sectors and the decision-making practices. In fact, the operational value of EDA has been strongly recognised especially within empirical or data-driven applications. Concerning the computation statistics, at the same time as the EDA birth also the earlier versions of statistical programming languages were released. Historically, the first example is the S[1] programming language for which some of the EDA pioneers - including J. K. Tukey - actively contributed to the design and development. S evolved in the premier statistical programming languages R[2]. Nowadays, a number of programming languages provide with statistical computing and graphics capabilities. Among these, Python[3] will be mainly considered and used later in this work. The increasing availability of tools to perform EDA is one of its main assets and allows for delivering interactive and reproducible computing that has terrifically changed the paradigm for the scientific discovery as well as for the operational business [65]. According to the above, we can objectively consider EDA one of the precursors as well as a pillar of the modern *Big Data* science.

[1] https://en.wikipedia.org/wiki/S_programming_language
[2] https://www.r-project.org
[3] https://www.python.org

2.1.2 Connection to the geospatial data analysis

Most of the concepts outlined above can be naturally conveyed into the context of geospatial data analysis. Indeed, the main difference between ESDA and EDA consists of the target data. Whereas EDA is designed to explore the characteristics of any generic series of observations (or a-spatial data), ESDA is specifically developed for dealing with geospatial data. Unlike a generic set of observations, geospatial data generally shows a strong dependence on the geographical location it refers to. This phenomenon is best described by the so-called *first law of Geography* which states that *"Everything is related to everything else, but near things are more related than distant things"* [103]. In the context of geospatial data analysis, this is known as spatial association, whose measurement helps to identify the degree of similarity between neighbouring observations in the geographical space, whether they are modelled as points, polygonal areas, or raster cells. Therefore, the spatial association matters because traditional statistical methods for data analysis (e.g. regressions) often rely on the hypothesis of independence for the observations. If the spatial association is present, this hypothesis is no longer valid, and the statistical inference might be biased or misinterpreted [76] [54].

According to the above statements, one of the main goals of ESDA is precisely to describe the spatial association and as well as to measure it by means of statistical tests. Considering the geographic nature of the problem, results are generally linked to maps. Thus, enabling a compelling visualisation of correlation patterns within a geospatial dataset. Here comes the affinity between ESDA and GIS software [5]. In fact, modern GIS software provides with reading, writing, numerical processing, and display capabilities to any digital geospatial data. This allows for developing all-in-one tools to perform ESDA and visually interact with both data and results.

To accommodate the needs of both statistical graphics and rigorous cartographic representations, specific visualisation techniques have been developed to accomplish ESDA tasks. Rigorous cartographic representations are characterised by defined and reproducible features which make maps metrics of the geographical space. Traditionally, these features are the scale, the topological relations, the projection, and the associated Coordinate Reference System (CRS). Statistical graphics are accounted by assigning a specific colour pattern or symbol style connected to the type of attribute to display. Categorical attributes are generally shown by varying symbols colour, hue, and shape as well as using labels. Quantitative data are portrayed using colour pattern spacing, hue, shape, and lightness variations (e.g. choropleth maps) [17]. In the context of ESDA, maps are no longer the final products they used to be in the traditional cartography. Indeed, maps function as storage of geospatial data analytics. Therefore, maps are considered a tool to gain insight into the distribution of data across locations as well as to discover data underlying structures, such as the spatial association [72]. Because most of ESDA methods are quantitative in nature, the choropleth map is one of the earliest adopted tools for ESDA purposes [104]. Depending on the specific application and on the data type, different classification intervals for the quantitative attributes can be adopted, e.g. the Natural Breaks, the Head/tail Breaks [68], and many others [91]. The selection of interval classification is crucial in quantitative mapping because of intuitive methods such as the equal interval classification do not properly capture underlying patterns for heavy-tailed or skewed attribute distributions [68]. Another popular ESDA mapping technique is the Cartogram [34] in which the scale of the map and its geometries are distorted to point out underlying data features better and guide the analyst in the exploration procedure. Nowadays, both choropleth maps and cartograms are largely employed for communication purposes rather than for analytics. Indeed, as a result of the link between ESDA and the computer

systems - such as GIS - a number of more powerful ESDA mapping tools have been made available [5]. These include interactive and dynamic desktop and web-based maps linking statistical graphics and embedding complex data analytics and browsing functionalities [6]. Both tools and mapping techniques are in a continuous evolution [114] that makes difficult to list and describe them all within this work. Valuable examples can be found scattered in the literature, e.g. [1], [93], and [97], as well as within most of the premier data visualization frameworks, libraries, and software [4,5,6,7,8].

I conclude this section by remarking the correct role of ESDA within a generic geospatial data analysis framework. What ESDA actually attempts to fill is the *gray area* between the raw data and the formal modelling as best summarised in Figure 2.1. With this in mind, results from ESDA have to be intended for testing assumptions, spotting anomalies, suggesting modelling strategies, and encouraging the analyst's interaction with both data and results rather than in a strict analytical or confirmatory sense.

2.2 Spatial association

A number of papers and books in the statistic literature spelt out the concept of spatial association at different levels of mathematical complexity, see, e.g. [24] [59], [2], and [39]. Historically, the spatial association has been appointed as *spatial autocorrelation*, among other terms. The spatial autocorrelation is a drift concept of the temporal autocorrelation as earlier introduced and under-

[4]http://datavizproject.com
[5]https://vega.github.io/vega/docs/data
[6]https://www.tidyverse.org
[7]https://plot.ly/python/maps
[8]https://geodacenter.github.io

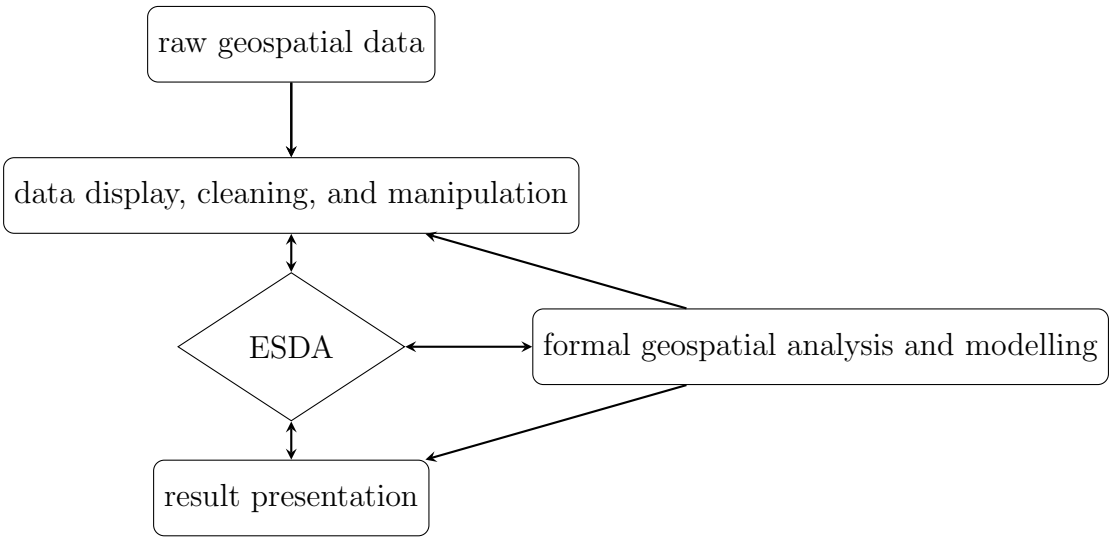


Figure 2.1: ESDA positioning in a generic geospatial data analysis framework.

stood in the time series analysis [113]. In this work, I prefer the use of the term spatial association because the autocorrelation generally refers to the analysis of a single variable whereas one of the goals here is to emphasise the extension of the concept to a multivariate context.

Section 2.1 focused on portraying the EDA and ESDA as well as some of their essential visualisation tools. To address the application of ESDA to the spatial association, the next section reports additional considerations on the spatial association and its implications into the geospatial analysis and the GIS science in general. The formulation of analytical methods to measure the spatial association - of interest for this work - is included in Chapter 3.

2.2.1 Why, when and how "location matters"

We broadly defined the spatial association as the measure of the degree at which similar things are also similarly arranged in space. A closer definition to the context of GIS science is derived from [55] and states that *"The spatial association represents the relationship between nearby spatial units, as seen on maps, where each unit is coded with a realisation of a variable"*. This relation-

ship characterises most of the geospatial data that is measured as a result of nonstationary or heterogeneous spatial processes [88] taking place in the real geographical space. Indeed, there are many instances in which an object location affects its behaviour. Snippets of evidence of the spatial association can be intuitively disclosed through simple examples. Housing prices are one among these. In fact, the location of a house will affect its selling price, and nearby houses are likely to be affected by the same neighbourhood effects. On the other way round, the selling price of a house can be only estimated by knowing its location together with, of course, its a-spatial features such as the building type. On one hand, this behaviour can be seen as a nuisance or a noise as it complicates statistical tests and the reliability of regression analyses by violating the independence assumption for the observations [39]. Moreover, collateral issues such as the analysis scale as well as the selection of the aggregation unit for the data might produce additional mis-specifications to the analysis [30]. On the other hand, once the structure of the spatial association is estimated, this information can be embedded into any prediction technique, e.g. the Kriging [26], thereby improving its accuracy. A fuller treatment of the topic can be found in [23], [76], [39], [54], and [31].

So far, we have outlined the effect of the spatial association mainly on the accuracy and reliability of traditional statistical methods. However, the spatial association provides with many uses and opportunities to any geospatial data analysis and particularly to the exploratory techniques. In the following list, which I partially amended from [55], the most meaningful applications of the spatial association concept are described.

- *Test of spatial stationarity.* Spatial stationarity implies that no significant variations in the spatial distribution of the analysys variable are present. Many spatial models require that spatial stationarity exists. Spatial association measures allow testing hypotheses of no spatial variation in dis-

tribution parameters such as the mean and the variance [63] [46].

- *Test on model mis-specification.* Properly specified spatial models require residuals, map onto the study region, to not be affected by any association between neighbour spatial units [23]. Tests of spatial independence for residuals are commonly performed using spatial association statistics [2].

- *Measure the strength of the location effects on any variable.* Spatial association coefficients help in weighing and understanding the strength of spatial effects in regression models [3] [44].

- *Investigate the influence of the spatial unit geometries on a variable and design spatial sampling.* Measures of spatial association change in known ways according to the configuration of spatial units. Spatial association measures help in understanding the role that spatial scale and aggregation strategies have on geospatial variables and their samples [61] [112] [115].

- *Test on hypotheses about spatial relationships and measure a spatial unit effect on other units (and vice versa).* Spatial association statistics are usually designed to assess the presence of a relationship among realisations of a single variable [4]. Tests may be extended to consider spatial relations between multiple variables [111] [8].

- *Identifying spatial clusters and outliers.* Spatial clusters and outliers are sub-regions of a spatial dataset which present significant strong similarity (or dissimilarity) for the observed variables. Spatial clustering algorithms rely on the conjecture of spatial association among some nearby values of one or more variables [87]. On the other way round, spatial association statistical and exploratory graphical tools allow for identifying spatial units that unduly influence or disturb spatial effects [4].

The list might be expanded whereas the most important concepts are there included. The last three points are of particular interest in this work. From an

exploratory perspective in GIS, the analysis of spatial association to point out interesting patterns, reciprocal influences, and significant effects characterising every spatial unit provides with plenty of chances to nimbly discover, by means ESDA mapping, relevant underlying characters of any geospatial datasets. Further considerations on this latter point are included in the next section.

2.2.2 Global versus local

Most of the long-established techniques aim at describing the spatial association in a dataset with a single measure, therefore gathering insights into its *global* behaviour. This is generally used for quantitative analysis with the purpose of assessing the presence and the degree of the spatial association among geospatial datasets as well as to integrate corrections for the spatial association into spatial modelling [37]. The assumption behind global methods is that spatial association properties are the same everywhere across the region of interest. This deficiency often masks spatial variations in the data by preventing analysts to detect inner pockets of spatial instability. As a consequence of the above, the development of *local* methods that account for inner spatial variations has been engaged in geography and connected disciplines in the recent past.

Spatial analysts have always been interested in local measures, that means to encode precisely both spatial characteristics and relationships of a particular site [55]. Moreover, the growing availability of geospatial data at a finer resolution as well as covering large areas, such as the satellite imagery, continent-wise road networks layers, etc. has uncovered the need for local methods [81]. This because the probability that regions with different properties would be encountered or considered within the modern spatial analysis has inevitably increased. Central to the topic is once again the need for exploring the spatial association. If the property of interest (e.g. precipitation, human population, traffic

jam, etc.) shows significantly different spatial association properties across the study region or across some scale of analysis, then e.g. a nonstationary model has to be considered for the analysis [81]. The same applies to the simultaneous analysis of multiple properties, i.e. in a multivariate context. In the case that different spatial processes partially overlap on the study region, the modelling phase requires to incorporate such a discovery [21]. By focusing o ESDA mapping, it comes naturally to prefer methods providing with local or mappable statistics whose representation and display are readily achievable with any GIS software.

A number of local methods have been developed in the context of geospatial data analysis. The logic behind local methods consists of allowing model parameters or statistical indicators vary as a function of the location. Typically, two approaches have been introduced to account for local variations in geospatial data modelling and analysis. These are summarised below.

- *Moving window methods*, whose model relationships between the data at the n locations closest to the centre of a moving spatial window. This enables to locally tune or fit parameter estimations thus allowing for local applications of global models that e.g. require for stationarity assumptions [29] [80].
- *Geographical weighting methods*, whose identify a neighbour relationship among the spatial realizations of a variable by means of a model matrix that can be directly embedded in the model formulation to account for local spatial variation in the data [45].

By considering the analysis of the spatial association, popular local methods are e.g. the Geographically Weighted Regression (GWR) [16], which allows regression parameters varying in space, and the Local Indicators of Spatial Association (LISA) [4] that allow evaluating the existence of spatial clusters in a

geospatial dataset. An exhaustive treatment of the local modelling in spatial analysis is provided by [81].

Relevant to this work are in particular the LISA. According to this, the next chapter is dedicated to the mathematical formulation and explanation of LISA by including also recent developments and extensions of these methods to a multivariate context.

Chapter 3

Methods for Measuring Spatial Association

The analysis of spatial association has challenged geographers and spatial statisticians over the last 70 years. Since the applications involving the spatial association vary considerably from field to field, many analysis methods have been created for different purposes. By confining the attention on GIS and mapping, the best-known methods are connected to or derived from a small family of statistics. These are the Moran's I [83] and the Geary's c [50] and the Getis-Ord G [51] statistics. All these statistics share a common general formulation and provide measures of spatial association that help to investigate patterns and interactions among spatially distributed variables. Originally, these statistics were designed as global measures of spatial association. Nevertheless, local versions have been proposed starting from the 90's [4] [51]. Actually, the G statistic, has been proposed directly in its local version and shows slightly different features than the I and the c. However, due to its relevance in mapping applications, the G is described together with the other two statistics in Section 3.3.

In Section 3.1, the general formulation of the above-mentioned spatial associa-

tion statistics is provided. This is followed by considerations on the significance testing for the selected spatial association statistics. A detailed description of the local versions of the selected statistics is the subject of Section 3.3.

3.1 General formulation

A number of methods for measuring spatial association have been proposed in the literature [39]. Nevertheless, all measures of spatial association can be traced back to a cross-product statistic as reported in Equation 3.1.

$$\Gamma = \sum_{i=1}^{n} \sum_{j=1}^{n} W_{i,j} R_{i,j} \; ; \; i \neq j \tag{3.1}$$

Where i and j is a couple of locations in the same geospatial dataset, with n equal to the total number of locations. The W_i,j matrix is the so-called contiguity or spatial weights matrix whose values define whether location j is a neighbour - or not - of location i (see e.g. Table 3.1). This is often expressed as a $n \times n$ symmetric matrix having values, e.g. ones, at i (i.e. row) and j (i.e. column) position if i and j are defined as neighbours, and zeros elsewhere. The term $R_{i,j}$ is a measure of the attribute similarity between locations i and j (see Section 3.3).

The neighbouring or contiguity relationship is specified by the analyst through rules such as a threshold distance for point data or edges and corners congruency for areal data. Most popular spatial weights types are summarised in Table 3.1. A fuller description of geographical weighting schemes can be found e.g. in [53] and [55].

The general formulation expressed by Equation 3.1 provides with a single mea-

Spatial Weight Type	**Description**
Distance-based	• Bandwidth $\bar{d}$ (distance = d): $W_{i,j}\begin{cases}1, 0 < d_{i,j} \leq \bar{d}\\ 0, d_{i,j} > \bar{d}\end{cases}$ • Power Distance (distance decay rate = α): $W_{i,j} = d_{i,j}^{-\alpha}$ • k-Nearest Neighbours (subset rule = N_k): $W_{i,j}\begin{cases}1, j \in N_k(i)\\ 0, j \notin N_k(i)\\ N_k(i) = \{d_{i,j^1} \leq d_{i,j^2} \leq \cdots \leq d_{i,j^k}\}\end{cases}$
Boundary-based	• Spatial Contiguity (edges/corners = bnd) $W_{i,j}\begin{cases}1, bnd(i) \cap bnd(j) \neq \oslash\\ 0, bnd(i) \cap bnd(j) = \oslash\end{cases}$

Table 3.1: Description of the most popular spatial weights types.

sure of spatial association for the whole geospatial dataset, thus describing its global behaviour. Generally speaking, this helps to identify both the degree and the type of spatial association. The degree is connected to the absolute value of the computed statistic. The type can be assessed by looking at the sign of the statistics or by comparing the computed value of the statistic to a reference one. Types of spatial association are basically two, namely the *positive* and the *negative* spatial association. Positive association stands for the significant presence of similar observations close in space, depicting a high spatial cluster activity for the geospatial dataset. Conversely, negative association implies a significant presence of dissimilar observations close in space, thus the likely presence of spatial outliers.

Despite the global spatial association measure provides with a compact and useful information for exploring characteristics of a geospatial dataset, it does not allow for presenting results directly on a map, that is of primary interest to GIS applications. To overcome this limitation, local versions for the spatial

association statistics have been introduced [4] [51]. Their general mathematical formulation is directly derived from the global statistic (Equation 3.1) by simply removing the summation on i, as shown in Equation 3.2.

$$\Gamma_i = \sum_{j=1}^{n} W_{i,j} R_{i,j} \; ; \; i \neq j \tag{3.2}$$

The latter leads to the definition of a single value for the statistic at each location i in the dataset, considering its j neighbours according to the adopted spatial weights matrix $W_{i,j}$. Therefore, these local versions are capable of suggesting the exact location of spatial clusters and outliers that is one among the most relevant tasks for GIS scientists.

3.2 Significance testing

A fundamental aspect connected to the use of local spatial association measures is the assessment of the significance of the computed statistics at each location. This implies computing reference or threshold values that allow to point out *interesting* locations, i.e. locations showing a strong dissimilarity with respect to the hypothesis of randomness, that is unlikely to be encountered across the whole study region when dealing with geospatial data. Indeed, the significance of the local spatial association statistics is inferred against the hypothesis of Complete Spatial Randomness (CSR). CSR aims at describing the distribution of the selected spatial association statistic in the case of a random arrangement of the observations within the n locations of the dataset.

The significance testing is key to the analysis of spatial association and ESDA in general. What it attains to demonstrate is - in fact - that the attribute val-

ues arrangement in some particular location of the dataset cannot be due to a random process. Therefore, there is likely a chance of modelling or explaining these non-random spatial patterns. This discovery is a crucial point for ESDA. A clear example of the above is reported in Figure 3.1. In the first map (Figure 3.1a), the actual spatial distribution of the African Americans population in the City of Chicago (USA) is portrayed. This variable shows an intense cluster activity thus providing insight into a well-known issue affecting the city, i.e. the racial segregation [22]. In Figure 3.1b, a random spatial distribution of the same variable is simulated, and the information on segregation remains unobserved. This clearly explains the central role played by the significance testing for the CSR hypothesis in the exploration of geospatial data.

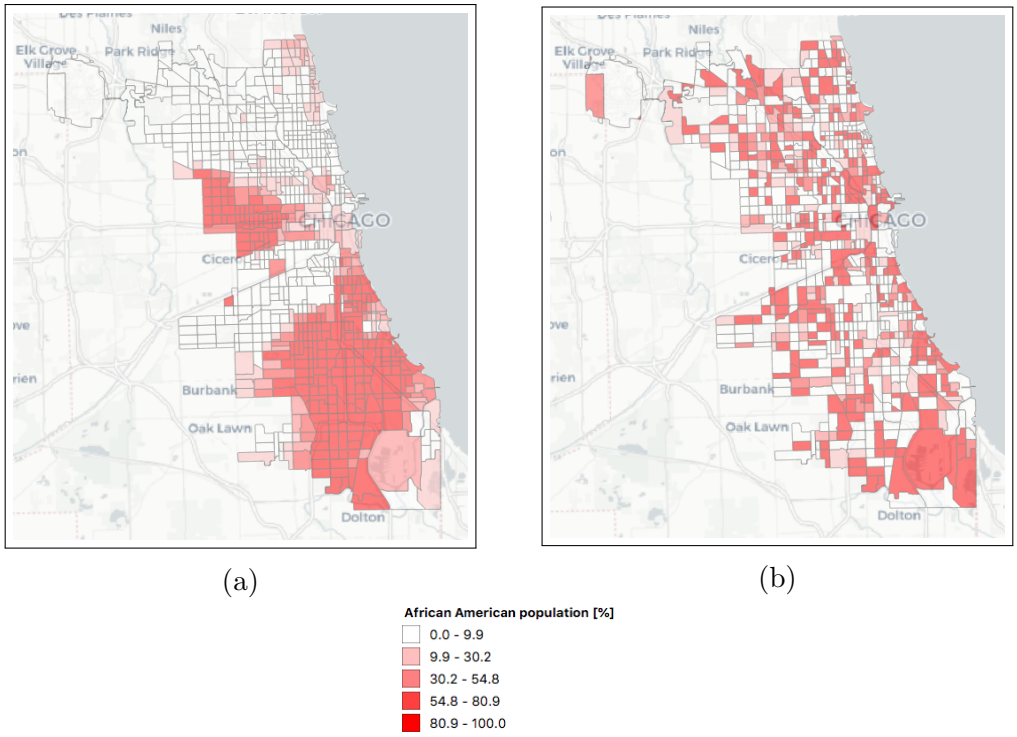


Figure 3.1: Share of African Americans population by city block in Chicago (USA) in the Year 2012. (**a**) the actual spatial distribution, and (**b**) a simulated random spatial distribution of the same variable. *Data source: http://robparal.blogspot.com. Basemap: ©OpenStreetMap contributors, ©CARTO.*

Analytical approximations to perform inference on local spatial association statistics have been discussed in the literature [98] [102]. The analytical-based inference is often used in the practice because less demanding in computational

terms. Analytical methods are based on the hypothesis of asymptotic normality for distributions of the local statistics [52]. However, the normality is shown under the conditions of the analytical testing of the ac CSR to poorly accurate procedures under some circumstances cite Leung e.g., for isolated locations of the dataset having fewer neighbours than the central ones cite Ord1995. Therefore, the use of computational methods is suggested to the practical inference. Computational methods are based on Monte Carlo like experiments such as the conditional permutations [12]. Practically, this approach consists of holding the value of the variable at location i fixed, random permute or shuffle the remaining values within the other n -1 location (Figure 3.1b) and recompute the local spatial association statistic. By repeating m times this process, an empiric reference distribution for the statistic under the CSR hypothesis is obtained for each location. Using this reference distribution, the pseudo p-value of the statistic at each location i can be computed according to Equation 3.3 [92].

$$p_i = \frac{b+1}{m+1} \tag{3.3}$$

Where m is the number of permutations and b is the number of times (out of m) that the statistics in the empirical reference distribution is equal or lower than the observed one. Smaller the pseudo p-value at a location i stronger the rejection of the null hypothesis (CSR). In turn, this means a higher probability that location i is *interesting* hence it belongs to a spatial cluster, or it is an outlier. A significance level α needs to be selected by the analyst to reject or accept the null hypothesis like any other statistical test.

The pseudo p-value estimated from the conditional permutations has to be cautiously interpreted [8]. This because it is not likely to properly reflect the

actual *Type I error*, that is the case when the null hypothesis (CSR) is true but is rejected thus a false positive is encountered [41]. Due to the computational procedure adopted in the conditional permutation approach, many of the values used to simulate a local measure of spatial association at a location i, are used again for a test on a neighbouring location by producing a large number of correlated tests. This ironically means that by searching for evidence of spatial association the tests are affected by the spatial association themselves [55]. In this unfavourable situation, the selected significance level or the p-values need to be adjusted to account for the *Type I error* with multiple comparisons. A number of empirical methods have been proposed [15] to control the error rate that is known as the False Discovery Rate (FDR). A graphical example of the above issue is included in Figure 3.2.

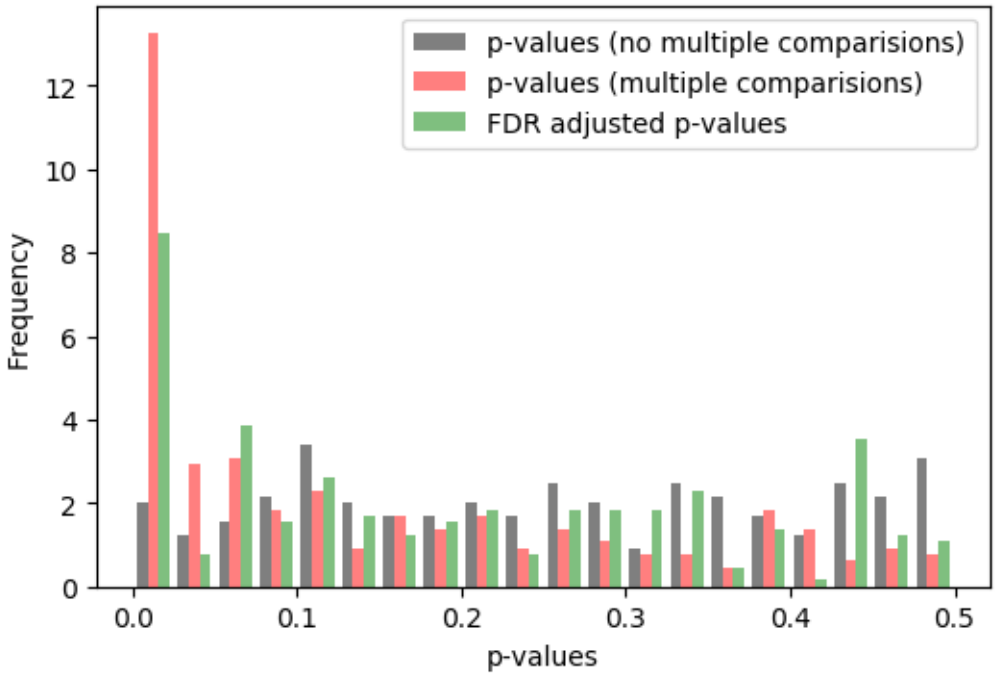


Figure 3.2: Schematic for the effect of correlated test on a p-values distribution. Black bins represent a simulated distribution from tests not affected by multiple comparisons, red bins a distribution in the presence of multiple comparisons, and green bins the affected distribution adjusted by means of FDR procedures.

Among these methods, the Benjamini-Hochberg procedure [10] is largely accepted as a standard to adjust p-values in a multiple comparisons problem.

The FDR is defined as the expected value of the ratio between the number of false positives in an experiment divided by the total number of discoveries in the experiment. This rate can be estimated from the pseudo p-values obtained from the conditional permutations and used to adjust the original p-values. The Benjamini-Hochberg procedure consists of a few steps as follows. First, the pseudo p-values for each observation p_{i} have to be sorted from smallest to largest and ranked such as $p_{\mathrm{i=1}} \leq p_{\mathrm{i=2}} \leq ... \leq p_{\mathrm{i=n}}$. Then, for a given significance level α , the larger rank i_{max} needs to be defined such as that the inequality expressed in Equation 3.4 is verified, where n is the number of observations. All observations with i lower or equal than i_{max} are then considered significant and considered to reject the null hypothesis.

$$p_{i,significant} \leq (\frac{i_{max}}{n})\alpha \tag{3.4}$$

An exhaustive technical discussion on the FDR procedure can be found in [41]. Due to the artefact introduced by the FDR, it is important to remind that the traditional definition of significance might not be appropriate to describe outcomes of the described inference procedure. That is the reason why many authors prefer the term *interesting* instead, which nevertheless best applies in the context of ESDA [8].

3.3 Local Indicators of Spatial Association (LISA)

The local versions of the spatial association statistics mentioned at the beginning of this chapter are known as Local Indicators of Spatial Association

of which LISA is the acronym. The mathematical formulation of LISA derives from a decomposition of their global parent statistic such as the sum of LISA computed for a geospatial dataset is proportional to the global statistic [4]. Along with this section, LISA are presented according to their enabling application into univariate, bivariate, and multivariate analysis contexts. The multivariate LISA description has to be intended as an experimental drift of the well-established LISA theory because of its recent development [8].

The possibility to automatically locate LISA outcomes on a map enables to depict spatial association patterns as well as to detect pockets of spatial association that cannot be discovered using global statistics. For these reasons, LISA mapping has become a cross-cutting practice within a number of disciplines ranging from ecology to economy, passing through health surveillance and land planning.

3.3.1 Univariate LISA

The Local Moran's I

Among LISA, the most popular is the Local Moran's I (Equation 3.5) [4], which is a local version of the Moran's I [83].

$$I_i = z_i \sum_{j=1}^{n} W_{i,j} z_j \; ; \; i \neq j \tag{3.5}$$

Where z_i and z_j are the standardized attribute values, such that their mean is zero and their variance is one, at locations i and j, and $W_{i,j}$ is the element of the spatial weights matrix. Generally, variables standardization is performed by means of Z-score scaling [36] as expressed in Equation 3.6, where x_i is an

unscaled observation of the considered variable, μ_x is the sample mean and σ_x the standard deviation.

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{3.6}$$

The link with the global parent statistics, the Global Moran's I [83], is that this is the average of the Local Moran's I computed at each location of the study region. A significant and positive value for I_i, expressed in deviation from the mean or Z-score, indicates that location i has similar values in its j neighbours, so it belongs to a spatial cluster. A significant and negative value indicates that location i has dissimilar values in its j neighbours, and it is, therefore, a spatial outlier. The significance is inferred considering the CSR hypothesis. Furthermore, by plotting on a Cartesian plane each couple (z_i, $W_{i,j}z_j$), where the $W_{i,j}z_j$ element is usually known as the *spatial lag* of the spatial attribute z_i, an additional information on the type of spatial association can be assessed. The resulting plot is known as the Moran Scatterplot [5]. Depending on the quadrant of the plane at which this couple is found, and belonging this couple to a significant cluster or outlier, the types of spatial association together with their descriptive conditions are described in the following list and by Equation 3.7:

- *Positive association of high values* (upper right quadrant)
- *Positive association of low values* (lower left quadrant)
- *Negative association of high z_i and low $W_{i,j}z_j$ values* (lower right quadrant)
- *Negative association of low z_i and high $W_{i,j}z_j$ values* (upper left quadrant).

$$location_i = \begin{cases} p_i \geq \alpha, & \text{not significant} \\ p_i < \alpha \wedge z_i \geq 0 \wedge W_{i,j} z_j \geq 0, & \text{high values cluster} \\ p_i < \alpha \wedge z_i < 0 \wedge W_{i,j} z_j < 0, & \text{low values cluster} \\ p_i < \alpha \wedge z_i \geq 0 \wedge W_{i,j} z_j < 0, & \text{high-low values outlier} \\ p_i < \alpha \wedge z_i < 0 \wedge W_{i,j} z_j \geq 0, & \text{low-high values outlier} \end{cases} \tag{3.7}$$

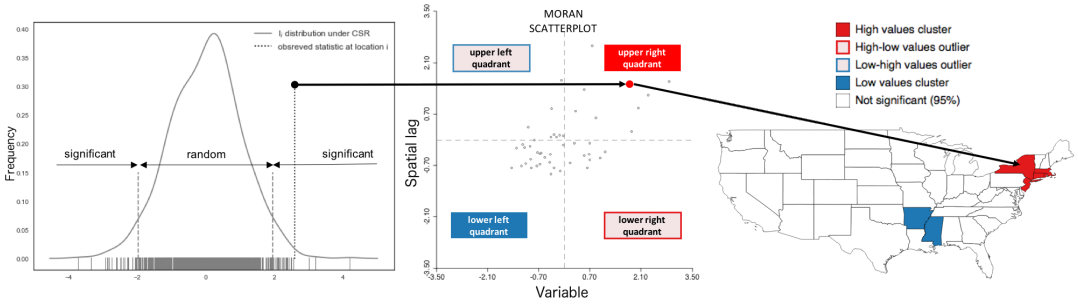


Figure 3.3: Schematic for the clusters and outliers mapping using the Local Moran's I.

The classification into quadrants is used to display results on a map by assigning a proper visualisation style to the geospatial dataset using GIS software (Figure 3.3).

The Local Geary's c

The Local Geary's c is the second LISA outlined in [4] and its formulation derives from the decomposition of the global parent statistic, the Geary's c proposed in the '50s by [50]. The Local Moran's I provides a local measure of spatial association as cross-products among a focal location and its neighbours. Differently, the Local Geary's c expresses the spatial association as a weighted average of squared distances in the attribute space between attribute values at a location i and that at each neighbouring location j, as shown in Equation 3.13.

$$c_i = \sum_{j=1}^{n} W_{i,j}(x_i - x_j)^2 \; ; i \neq j \tag{3.8}$$

Where x_i and x_j are the attribute values at locations i and j, and $W_{i,j}$ is the element of the spatial weights matrix. The significance is inferred considering the CSR hypothesis. Due to the analytical formulation, this statistics focuses on dissimilarity rather than correlation. Therefore, a significant and less than the mean value for c_i depicts similarity, hence positive spatial association. A significant c_i value greater than its expected value indicates dissimilarity or negative spatial association. On the other way round, a significant c_i value lower than its expected value indicates similarity or positive spatial association. The expected value $\mathbb{E}[c]$ can be either computed analytically, e.g. the sample mean, or from an empirical reference distribution derived by the conditional permutations used for significance testing, as for the other statistics explained before. These conditions are specified by Equation 3.9 and in Figure 3.4.

$$location_i = \begin{cases} p_i \geq \alpha, & \text{not significant} \\ p_i < \alpha \wedge c_i < \mathbb{E}[c], & \text{cluster} \\ p_i < \alpha \wedge c_i \geq \mathbb{E}[c], & \text{outlier} \end{cases} \tag{3.9}$$

The Local Geary's c does not provide any chance of classifying detected spatial cluster and outliers such as for the Local Moran's I. A partial classification can be achieved by the simultaneous analysis of resulting cluster maps form the Local Geary's c and the Local Moran's I by using dynamic linking and brushing visual techniques [6]. Due to this limitation, other LISA are usually preferred for map-based researches. However, the Local Geary's c can be extended to the

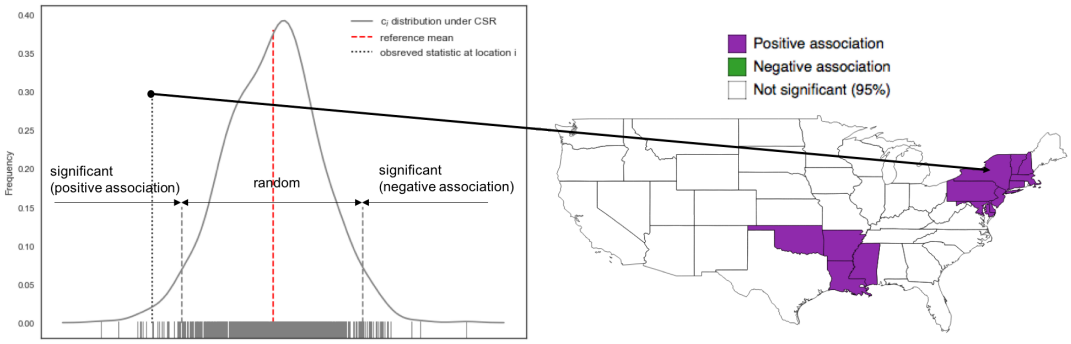


Figure 3.4: Schematic for the spatial association mapping using the Local Geary's c.

analysis of multiple attributes, thus to a multivariate context, making it central to this research work. The multivariate extension of this LISA is presented later in this section.

The Local Getis-Ord G

The last univariate statistic here presented is the Local Getis-Ord G [51]. This is not a LISA in a formal sense because there is not a direct connection to a global parent statistic. Nevertheless, the Local Getis-Ord G is likely the most popular techniques for detecting spatial clusters. This statistic provides a measure of spatial association by comparing the sum of the values x_j in the neighbourhood of the focal location i with the sum of all locations. The significance is inferred, once again, considering the CSR hypothesis. Two versions of the Local Getis-Ord G are available as reported in Equation 3.10. The G_i that evaluates the spatial association as values similarity in the proximity of the focal location i, and the G_i^* allowing for clustering detection such as the location i is included in the sum by introducing non-zero diagonal values in the spatial wights matrix W.

$$G_i^* = \frac{\sum_{j=1}^{n} W_{i,j} x_j}{\sum_{j=1}^{n} x_j} \;\; ; \;\; i \neq j; \;\; {}^*i \in j \tag{3.10}$$

Conversely to the Local Moran's I and the Local Geary's c, the interpretation of the results is straightforward. Significant and positive values, expressed as deviations from the mean or Z-scores ($\mathbb{Z}[G_i^*]$), depict clusters of high values or hot spots, and vice versa significant and negative values clusters of low values or cold spots as expressed by Equation 3.11 and Figure 3.5. The negative spatial association, i.e. spatial outliers, cannot be spotted by using the Local Getis-Ord G.

$$location_i = \begin{cases} p_i \geq \alpha, & \text{not significant} \\ p_i < \alpha \wedge \mathbb{Z}[G_i^*] \geq 0, & \text{hot spot} \\ p_i < \alpha \wedge \mathbb{Z}[G_i^*] < 0, & \text{cold spot} \end{cases} \tag{3.11}$$

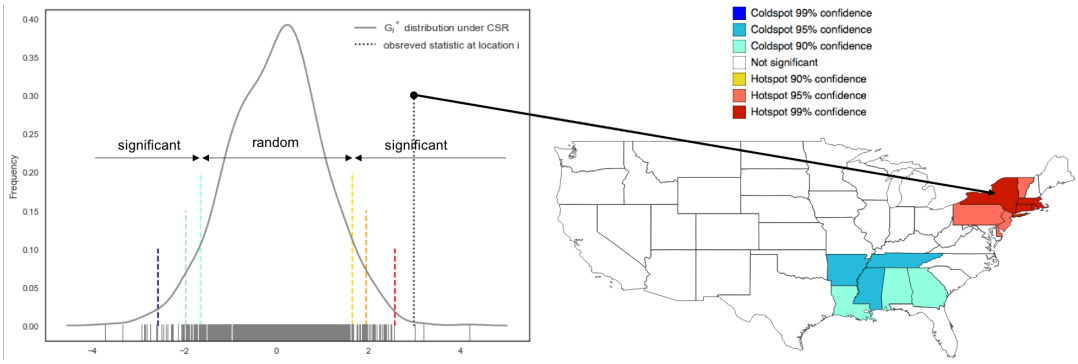


Figure 3.5: Schematic for the hot and cold spots mapping using the Local Getis-Ord G with multiple significance levels.

Univariate LISA statistics provide analysts with different rationales for investigating the spatial association. These include similarity measures based on

additive processes (G), cross-products (I), and squared distances (c). Combinations of these indicators are generally considered during the exploitative phase or mapping of the spatial association in a dataset, hence increasing the inference power against alternative hypotheses. This means that the rejection of the null hypothesis (CSR) can be achieved with high confidence while alternatives remain unknown. Indeed, spatial clusters and outliers are suggested by LISA but not explained. LISA do not provide insights into the possible spatial processes generating the patterns. Univariate LISA results might also suffer from underlying interaction between the analysis variable and other covariates which might act at different scales or just on portions of the study region. While the explanation of both univariate and multivariate spatial processes remains out of LISA purposes, the exploration of multivariate interactions in terms of spatial association can be achieved by LISA extensions to the multivariate context. This topic is presented in the following.

3.3.2 Bivariate LISA

The Local Moran's I is suitable for both univariate and bivariate analyses. In the bivariate setting, the geospatial dataset has two attributes measured at each location. The bivariate Local Moran I is introduced in Equation 3.12 to enable a joint spatial association analysis of the two attributes [111] [75].

$$I^i_{k,l} = z^i_k \sum_{j=1}^{n} W_{i,j} z^j_l \; ; \; i \neq j \tag{3.12}$$

Conversely to Equation 3.5, the z_i is substituted with the standardized values of the base attribute z^i_k, and z_j with the standardized values of a second attribute in the neighbours locations z^j_l. Clearly, the standardization procedure (Equation

3.6) becomes here necessary in order to scale and compare the two different attributes. By doing so, a significant and positive value for $I^i_{k,l}$ indicates that location i has the value of the first attribute similar with values of the second attribute in its j neighbours. Location i is - therefore - a joint spatial cluster for the two attributes. A significant and negative value for $I^i_{k,l}$ hence location i being a joint spatial outlier for the two attributes. Results can be linked and mapped using the quadrant classification of the Moran Scatterplot, as for the univariate version. Applications of the bivariate Local Moran's I are illustrated in Chapter 5. The Local Geary's c could be adopted in a bivariate context. I preferred to report in the following its general extension to the multivariate case which naturally applies also to the bivariate analysis.

3.3.3 Multivariate LISA

In a multivariate context, i.e. when k attributes are measured at each location and their spatial association needs to be investigated simultaneously, no consolidated methods are available in the literature. Recently, an extension of LISA has been finally proposed by [8], potentially enabling multivariate spatial association analyses and mapping. This is an extension of the Local Geary's c statistic [4]. This LISA was originally designed to accomplish univariate analysis, and it is a measure of the squared distance - in the attribute space - between the attribute value at a location i and that at each neighbouring location j. The multivariate extension of the Local Geary's c collapses the k squared distances into a weighted sum by providing a single value for the statistic at each location of a geospatial dataset, as shown in Equation 3.13.

$$c_{k,i} = \sum_{v=1}^{k} \sum_{j=1}^{n} W_{i,j} d^2_{v_{i,j}} \; ; \; d^2_{v_{i,j}} = (z_{v,i} - z_{v,j})^2; \; i \neq j \tag{3.13}$$

Where $d^2_{v_{i,j}}$ is the k-dimensional squared distance, in the attribute space, between the standardize z_v attribute values at locations i and j, and $W_{i,j}$ is the element of the spatial weights matrix. Even more than in the bivariate analysis, the standardization procedure (Equation 3.6) is here crucial for comparing the k different attributes. As for the univariate case, the interpretation of spatial association using the multivariate Local Geary's c is less intuitive as that of the other LISA such as the Local Moran's I. A significant value of $c_{k,i}$ that is less than its expected value suggests a positive spatial association. While a significant and higher value suggests negative spatial association [8]. The expected value can be computed analytically or from an empirical reference distribution derived by the conditional permutations, as for the univariate version of the statistic. Cluster and outliers classification, such as the one introduced with the Moran Scatterplot, is not achievable in this case, leading to a strong limitation of exploring the spatial association by the exclusive use of the multivariate Local Geary's c. However, *interesting* locations in the dataset can be identified and later analysed exploiting other data analysis techniques. This topic is the subject of Chapter 6.

Being recently developed, the multivariate extension of the Local Geary's c is neither exhaustively tested nor fully integrated into stable software tools. For these reasons, a proper interpretation of the outcomes cannot be currently performed in an objective manner. In Chapter 6, we attempt to an application of this new LISA for demonstrating the technical feasibility of the methodology, its integration into a GIS environment, and its consolidation with traditional EDA methods, rather than for confirming its advantages and the practical implications of the results for the selected case study.

Chapter 4

Software Tools

The strong connection with ESDA and graphical techniques has naturally led to the spread of software tools dedicated to geospatial data exploration by means of statistical mapping. This has been significantly facilitated by the diffusion of GIS and the terrific improvements that both computer graphics and computational power faced along the last two decades. Nowadays, ESDA and - in particular - LISA rely on many software implementations available as libraries for many programming languages, statistical software, and GIS software modules. Some of these tools are released with open licenses[1] thus allowing users free access to both software functionalities and the source code. This provides analysts with plenty of tools and opportunities for investigating, testing, and extending spatial association methods by also empowering geospatial data analysis and modelling with dynamic data interactions. Focusing on spatial association testing and mapping, some of the most popular and cutting-edge tools are introduced in the following list.

- *GeoDa*[2] is a FOSS distributed under the GNU General Public License (GPL)[3] that provides geospatial data analysis and visualization function-

[1] https://opensource.org/licenses
[2] https://geodacenter.github.io
[3] https://www.gnu.org/licenses/gpl.html

alities by means of an interactive GUI to explore and modelling spatial patterns. The software runs on the main Operative Systems (OS) like Windows, macOS and Ubuntu. The analytical functionalities are implemented as a collection of C++ classes with associated methods. The analytics include spatial association and spatial regression models as well as EDA and mapping tools. GeoDa supports a large variety of vector data formats such as shapefiles, geodatabases, GeoJSON, and Keyhole Markup Language (KML) layers thanks to the integration with the Geospatial Data Abstraction Library (GDAL)[4]. Basemaps can be added to any map view to improve data and results visualisation. GeoDa provides extensive support for LISA mapping as well as multivariate EDA by means of linking and brushing operations between maps and statistical graphs (Figure 4.1).

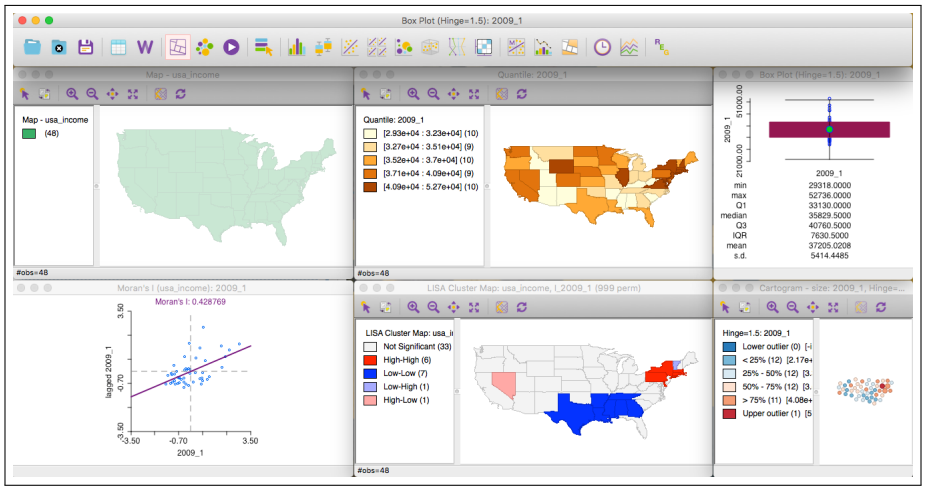


Figure 4.1: GeoDa dynamic interfaces including statistical maps and linked EDA graphs.

- *The ArcGIS ESDA extension*[5] is part of the Spatial Statistics Toolbox of the premier proprietary desktop GIS software ESRI® ArcGIS[6]. This software runs on Windows OS only whereas a cloud-based version, the ArcGIS

[4]https://www.gdal.org
[5]https://tinyurl.com/yah6f7cx
[6]http://desktop.arcgis.com

Online[7], has been recently made available. ArcGIS allows user for creating custom desktop and Web applications by serving a number of Application Programming Interfaces (API) for many programming languages such as Python, JavaScript, and Java. The ESDA extension provides with a set of techniques for describing and modelling geospatial data including EDA statistical graphs, LISA mapping (Figure 4.2), and many global measures of spatial association. These are fully integrated into a complete GIS environment that eases any geospatial data manipulation and display through a user-friendly GUI.

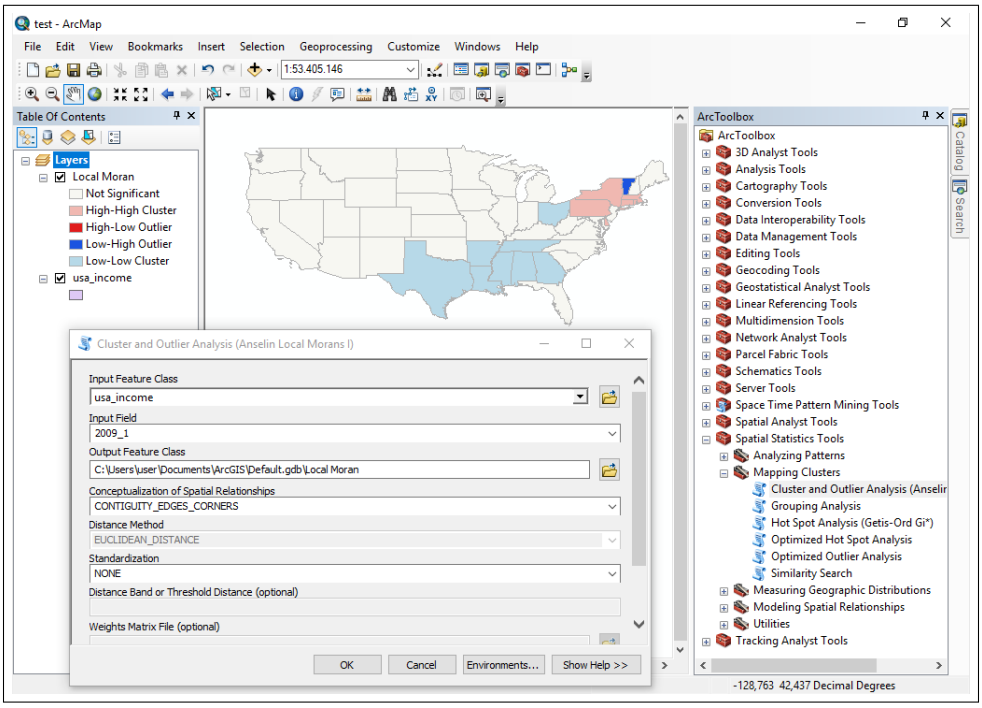


Figure 4.2: LISA mapping with ESRI® ArcGIS.

- *The Spatial Dependence: Weighting Schemes, Statistics and Models (spdep)* [8] is a cross-platform package for the statistical programming language R dedicated to geospatial data analysis. It is a FOSS released with the GPL license that includes a collection of functions to perform a number of spatial analytics such as pattern analysis, spatial regressions and

[7]https://www.esri.com/en-us/arcgis/products/arcgis-online
[8]https://cran.r-project.org/web/packages/spdep

tests for spatial association. Thanks to the capabilities of the R language, the spdep allows for integrating with other packages by extending both supported data formats and mapping capabilities (Figure 4.3). The access to functionalities is enabled only through the command line through scripting operations.

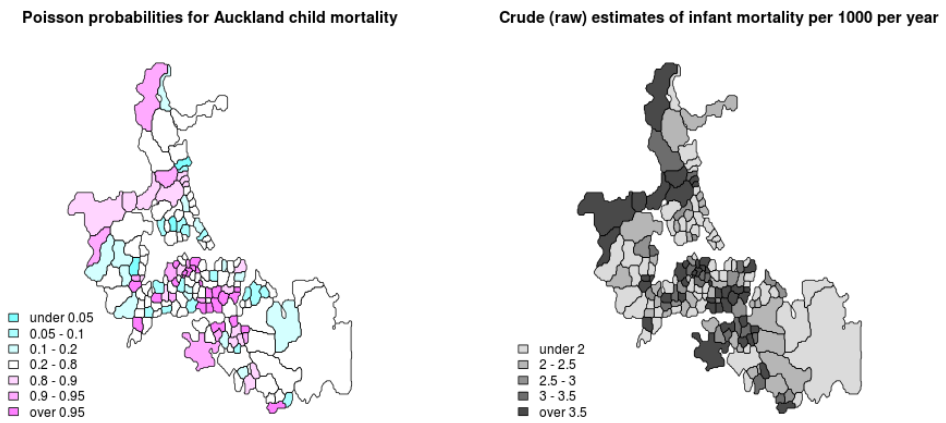
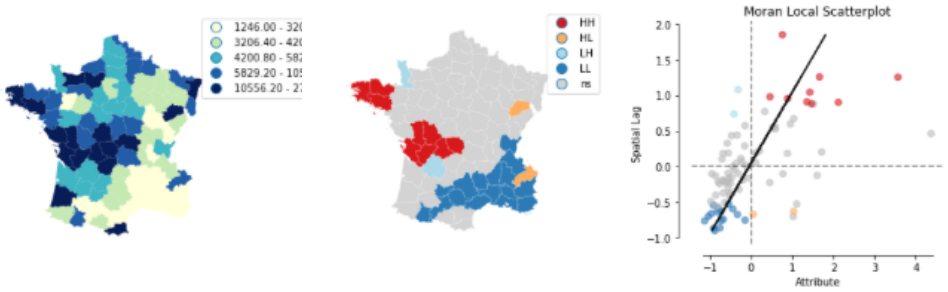


Figure 4.3: Sample statistical maps computed using the spdep R package. *Source: https://rdrr.io/rforge/spdep/man/probmap.html*.

- *The Python Spatial Analysis Library (PySAL)*[9] is a FOSS cross-platform library for geospatial data science data written in Python and release under the 3-Clause Berkeley Software Distribution (BSD) license[10]. PySAL project was started with the goal of leveraging existing software tools development such as GeoDa and Space-Time Analysis of Regional Systems (STARS)[11] in order to supply a core of analytical functions to support and extend spatial analysis applications. This also by considering the emerging role of Python as the scripting language for geospatial analysis since its adoption by ESRI®, the leading commercial GIS software provider, as well as its growing popularity among spatial analysts. PySAL consists of multiple packages that implement a number of spatial statistics, spatial association tests, and spatial regression models. It also provides

[9] http://pysal.org
[10] https://opensource.org/licenses/BSD-3-Clause
[11] http://regionalanalysislab.org/index.php/Main/STARS

built-in methods to display and visually analyse spatial datasets and spatial statistic outputs by means of ESDA graphic techniques (Figure 4.4). Each package is deployed as a stand-alone component that eases the integration into data analysis frameworks by optimising the dependency requirements as well as facilitates the development and maintenance of the library. PySAL supports multiple vector data formats such as shapefiles, and text-based file formats like Comma-separated Values (CSV), GenePix Array List (GAL) files, and DataBase File (DBF). Being a programming library, the access to functionalities is enabled exclusively through the command line. Only experimental GUIs to exploit the PySAL library are currently available, the stand-alone FOSS CAST[12] is an example.



(a) (b) (c)

Figure 4.4: An example of the PySAL analytic and graphic functionalities. A choropleth map for a spatial variable (**a**), the local Moran's I map computed from the variable (**b**), and the corresponding Moran Scatterplot (**c**). *Source: https://github.com/pysal/splot.*

Modules for spatial association analysis can also be found in different statistical software such as SAS©,[13], some FOSS GIS like SAGA-GIS[14], or embedded into Web-based analytical platforms such as CARTO©,[15]. Nevertheless, the tools included in the previous list represent the state-of-art concerning LISA mapping that is the central topic of this work. In particular, PySAL is considered in the next section to highlight the opportunities enabled by FOSS of developing

[12]https://geodacenter.github.io/CAST
[13]https://support.sas.com/rnd/app/stat/procedures/SpatialAnalysis.html
[14]http://saga-gis.org
[15]https://carto.com/platform/spatial-data-science/

custom LISA mapping applications and sharing them to the users' community.

4.1 The Hotspot Anlysis Plugin for QGIS

Within FOSS GIS, there is still a lack of functionality dedicated to spatial association statistics and modelling [7]. However, the nature of the different FOSS projects facilitates their extension and integration with other software tools, thus enabling a valuable mix of custom geospatial data analysis tools with a wide range of GIS mapping functionalities. This is the case of QGIS[16] that is currently recognized as one of the leading FOSS GIS. QGIS is a user-friendly FOSS GIS licensed under the GPL that provides geospatial data manipulation and analysis tools both on desktop and Web environments. Being a FOSS, users are allowed to contribute the source code and the maintenance of the software. QGIS is an official project of the Open Source Geospatial Foundation (OSGeo) that provides with support, and coordination to the worldwide community of FOSS GIS developers. The software runs on most Unix platforms, Windows, macOS, and Android OS and supports numerous vector, raster, and database data formats. QGIS is developed in C++ while the GUI is based on the Qt[17] framework. Actually, QGIS is much more than a GIS software. It provides a complete development environment that includes extensive support for Python scripting and programming [82]. Custom scripts and data processing pipelines can be run in QGIS by exploiting the Python QGIS (PyQGIS) APIs [74] as well as by coupling them with functionalities from both external programming languages such as R, and FOSS GIS software like GRASS GIS[18], SAGA-GIS, the Sentinel Application Platform (SNAP)[19], and many others [60]. The feature that mainly characterises QGIS is, however, the enabling possibility for users

[16]https://www.qgis.org
[17]https://www.qt.io
[18]https://grass.osgeo.org
[19]http://step.esa.int/main/toolboxes/snap/

to develop and publish custom plugins. QGIS Plugins[20] are independently developed software packages that can be installed and run on QGIS to extend its core functionalities. QGIS Plugins are written in Python according to a common framework that provides the minimum requirements for a plugin to work. The interoperability between plugins and the core QGIS software is provided by the PyQGIS APIs while the plugin interfaces are developed using PyQt[21] which is one of the most popular Python bindings for the Qt framework. QGIS provides an official share repository to which any user can download available plugins or deploy their ones. Generally, the plugins development and is carried out using Web-based Version Control Systems (VCS) such as the GitHub[22], which is the VCS used also for the QGIS core development. Plugins can embed functionalities derived from Python libraries that are included in the default QGIS Python installation, by providing documentation on the dependencies requirements as well as the licensing compatibility of the external resources adopted and QGIS. This aspect is important, especially for published plugins. In fact, a plugin can be produced for personal use only or shared onto the official repository where a preliminary check of the code is performed by the repository managers. Some of the top-ranking plugins, initially developed by users, have been later included in the QGIS core, thus improving the default software capabilities during the time. This virtuous development strategy, perhaps characterizing most of the FOSS projects, represents the main strength of QGIS and, at the same time, is one of the main reason for its popularity. Due to the above-described characteristics and opportunities, I selected QGIS as a platform for developing new FOSS GIS tools dedicated to LISA mapping that is one of the outputs of the presented work.

Considering leading FOSS GIS like QGIS, no core ESDA functionalities imple-

[20] https://plugins.qgis.org
[21] https://wiki.python.org/moin/PyQt
[22] https://github.com/

menting LISA mapping have been made available so far. These spatial statistics can be accessed only through command line thus without the support of a dedicated GUI. In order to provide QGIS users with ESDA functionalities, we develop an experimental plugin called Hotspot Analysis [89]. This allows performing LISA computations such as the Local Getis-Ord G_i^* and Local Moran's I, both in its univariate and bivariate versions, on a geospatial dataset and automatically display results on a map as shown in Figure 4.5. The plugin is written in Python and exploits some functionalities provided by the *PySAL* ESDA module[23], coupled with some of the digital mapping facilities enabled by the PyQGIS APIs. The interface is built using on PyQt like any other QGIS plugin. Both the source code and the user documentation are available on the GitHub[24]. The stable version is also published on the official QGIS plugins web portal[25] which can be accessed and installed directly from the QGIS software interface through the plugin manager menu. A description of the plugin functionalities, derived from a recent paper that I co-authored [89], is reported in the following.

The Hotspot Analysis Plugin requires as input a shapefile of polygons (Figure 4.5) or points (Figure 4.6) with associated a projected CRS and - at least - one numerical attribute associated at each location. The attribute is the spatial realisation of the variable for which a LISA is computed. This information has to be assigned to pointwise locations or parcels covering the area under investigation, e.g. city block centroids, pixels of a regular grid, etc. Spatial relationships between neighbour geometries are considered by creating a spatial weights binary matrix, exploiting the dedicated PySAL functionality[26]. A limited set of methods for creating the spatial weights matrix is currently available in the plugin. Concerning points, the default spatial weights matrix is created

[23]https://pysal.readthedocs.io/en/latest/library/esda
[24]https://github.com/danioxoli/HotSpotAnalysis_Plugin
[25]https://plugins.qgis.org/plugins/HotspotAnalysis
[26]https://pysal.readthedocs.io/en/latest/library/weights

using a fixed distance band, expressed with the same unit of measure of the projected CRS of the input shapefile. Alternatively, the matrix can be created using the K-Nearest Neighbours (KNN) approach, which enables to define a relation for each point of the dataset with its K nearest points, where K value is set by the user. For polygons, a first order *Queen's case* contiguity matrix is used, i.e. edges and/or corners contiguity. The output layer is displayed with an automatic style that combines Z-scores and p-values allowing an intuitive visualisation of the detected local spatial clusters. For what it concerns the Getis-Ord G_i^*, a positive and statistically significant Z-score indicates a cluster of high values (hot spot). A negative and statistically significant Z-score indicates a cluster of low values (cold spot). Concerning the Local Moran's I and the Local Moran Bivariate, statistically significant Z-scores are translated into quadrant values (q-values) of the Moran Scatterplot which are included into a separate column of the attribute table. The q-values depict the presence of clusters or outliers within the dataset. Plugin additional functionalities are the following. The selection of row standardized spatial weights matrix, instead of the default binary version. Row standardization consists of dividing each matrix element by the sum of the row to which the element belongs to. This normalization is traditionally used to prevent LISA begin biased by an uneven neighbours availability at isolated or marginal locations in the dataset [71]. A second option is the computation of statistical significance using the permutation approach instead of normality approximation, and lastly the optimisation of the fixed distance band selection. The optimisation consits of defining a range of possible distances to be tested by the plugin. The optimal distance that is used by the plugin for computations is the one maximising the Z-score of the Global Morans I [83] for the dataset. This distance reflects in principle where the spatial process that promotes clustering is most pronounced[27]. The default distance band suggested for any input shapefile of points is the mini-

[27]https://tinyurl.com/y8d4y3t7

mum distance that guarantees at least one neighbour to each point. The plugin functionalities are summarized in Table A.1 included in Appendix A.

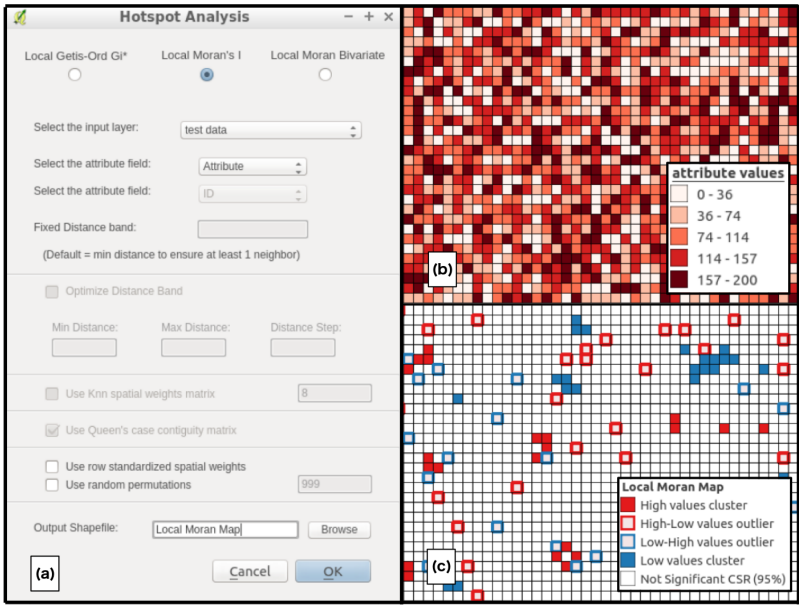


Figure 4.5: The GUI of the Hotspot Analysis Plugin (**a**), a sample input data (polygons grid) with an attribute distributed in space (**b**), and (**c**) an example of the LISA output map from the Local Moran's I computation.

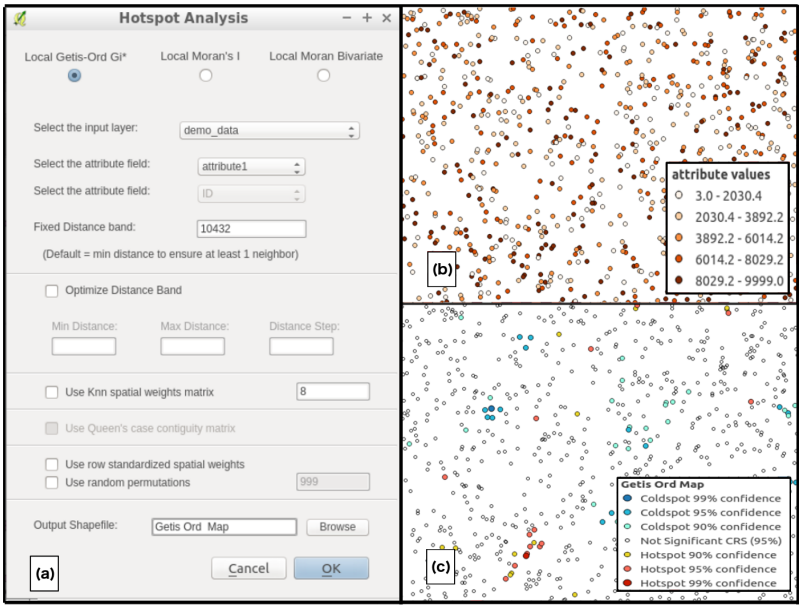


Figure 4.6: The GUI of the Hotspot Analysis Plugin (**a**), a sample input data (points layer) with an attribute distributed in space (**b**), and (**c**) an example of the LISA output map from the Local Getis-Ord G_i^* computation.

In Chapter 5, sample applications of the plugin to spatial association analysis on real geospatial datasets are presented. Regarding the multivariate LISA,

only an early module of the GeoDA software has been currently released. The module provides with simple features to compute multivariate Local Geary's c maps from many geospatial data formats. However, advanced options for the post-processing of the output maps are not available within the software. For testing the multivariate Local Geary's c, custom Python scripts are therefore developed and used in this work enabling full control of both the outputs and the mapping experiment settings. An extensive discussion of the multivariate LISA computations is included in Chapter 6.

Chapter 5

Case Study Applications

In this Chapter, LISA applications on real data are illustrated with the aim of helping the readers in better understanding the concepts of chapters 2 and 3. At the same time, the practical use ESDA software is described. The Hotspot Analysis Plugin is considered for this purpose (Section 4.1). Selected case studies focus on both univariate and bivariate LISA mapping. The complete case studies discussion is published in recent papers that I co-authored [11][89][90], from which part of the following is derived.

5.1 Univariate LISA mapping

5.1.1 Case study: Slow mobility spatial patterns

Nowadays, there is an emerging use of geospatial data and location-based services not only among professionals but also by lay users during leisure activities. A relevant example is the use of mobile devices for tracking outdoor activities such as running, hiking, cycling, ski touring, etc. The information produced in this context is then often shared by users on dedicated Web portals allowing for visualizing, downloading, and reusing the content among the users' commu-

nity. A popular content which can be found on these platforms, in addition to pictures and textual descriptions of routes and itineraries, consist of Global Navigation Satellite Systems (GNSS) tracking data depicting the real users' activity along territories [18]. This example best describes the underlying role of geospatial data for recreational activities. The activities mentioned above are often referred to as *slow mobility* [49]. This has been increasingly recognized as an asset for the sustainable development of the territories [73] because of its low environmental impact as well as positive effects on the citizens' health. Due to the geospatial nature characterising most of the user-generated content portraying slow mobility activities, ESDA can be adopted as a tool for investigating the spatial characters of this phenomenon. In the following, an example of univariate LISA mapping applied to slow mobility is reported.

As a case study, the Lombardy Region (Northern Italy) was selected, which includes different landscapes as well as environments ranging from highly populated cities to the alpine glaciers, passing through its famous subalpine lakes and vast plains (Figure 6.2a). Thanks to this territorial variety, the Lombardy Region is a good candidate for investigating slow mobility through its wide range of different environments that favour the practice of these kinds of activities. Considered data consists of user-generated GNSS tracks retrieved from the Wikiloc platform[1]. This is a free Web-based service offering users the possibility of sharing their outdoor experiences using GNSS trails as well as access to the platform content through advanced map-based functionalities. The available GNSS tracks for the Lombardy Region were downloaded in GPS eXchange Format (GPX). This format provides linear layers constituted by chronologically ordered waypoints (point coordinates with time stamps) enabling full tracking of movements registered by a GNSS device. Data collection referred to the period January - March 2016.

[1]https://www.wikiloc.com

The collected GPX files were stored in a PostgreSQL-PostGIS[2] database. A preliminary data cleaning procedure was applied to filter out trails with associated average speed, computed from the correspondent GPX tracks, greater than 22 [km/h]. According to [56], this threshold would cover mostly non-motorised transportation, which was present inside the Wikiloc database in some proportion. GPX tracks were re-sample at a 15 [s] time gap to account for the possible differences in terms of waypoints sampling by the different users' devices. The waypoints considered were about 2.100.000. The dataset was split into two groups according to the time when each track was generated, namely weekdays and weekends. By doing so, it was possible to introduce a temporal dimension into the analysis. Considering slow mobility patterns, this temporal distinction is interesting for outlining users interactions with the territory that might show important variations between the two periods. Finally, the waypoints counts within each municipality area of the Lombardy Region was performed for the two considered periods. The count values were associated with the centroid of the corresponding municipality and represented the variable on which LISA mapping was performed. Selected LISA was the Local Getis-Ord G_i^*. Hot spot and cold spot maps were created using the Hotspot Analysis Plugin for QGIS (see Chapter 4) using a row-standardized spatial weights matrix computed by a fixed-distance band weighting schema.

By mean of LISA maps, it is possible to discover clusters, at a regional scale, of popular destinations for slow mobility activities. This reflects the actual users' activities along territories whose attractiveness is explored by means of spatial association among the registered waypoints counts. Results are reported in Figure 5.1. During weekdays, hot spots concentrated mainly around some of the main cities (i.e. Milano and Brescia) as well as in the Alpine area (Figure

[2]https://postgis.net

5.1b). During weekends, a large hot spots concentration appears all along the Subalpine area (Figure 5.1c), while cold spots clustered mainly in the Plain area. With the Wikiloc data, the Local Getis-Ord G_i^* highlights the different concentration of activities which is explored accounting for the territorial features of Lombardy Region. This experiment provides insight into a large-scale phenomenon, such as the slow mobility, through simple maps that help in better discovering hidden patterns in the data which is among the main goals of LISA.

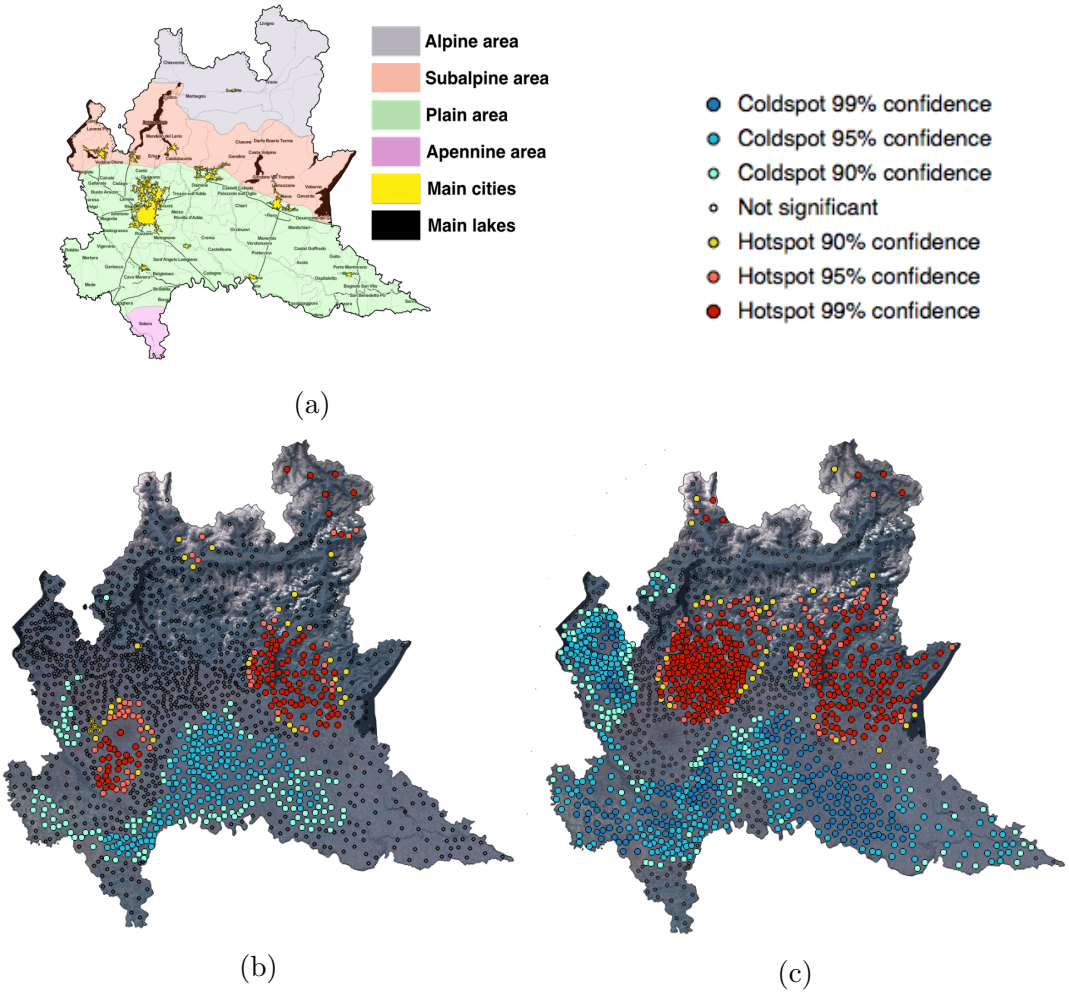


Figure 5.1: (**a**) reference map including the main territorial features of the Lombardy Region. Local Getis-Ord G_i^* maps for the Wikiloc waypoints counts per municiaplity during weekdays (**b**) and weekends (**c**). *Source (**a**): [11], Basemap: (**a**) ©OpenStreetMap contributors, ©Stamen Design, (**b**,**c**)* ©MapQuest.

5.2 Bivariate LISA mapping

In this section, two sample applications of the bivariate LISA mapping are reported. The first application is connected to the spatial association analysis between users' feedbacks and the prices for Airbnb©,[3] accommodations in the City of Venice (Northern Italy). The second application deal with the spatial association analysis of the soil consumed and the average income per capita in Italy.

5.2.1 Case study: Tourist accommodations analysis

In the first example, the *QGIS* and the Hotspot Analysis Plugin is used to perform LISA mapping on the Airbnb© accommodations. Airbnb© is a Web-based marketplace and hospitality service, enabling people to lease or rent lodgings for vacations or short-term staying. It relies on a large and widespread community of user worldwide and, among its services, it provides a collection/sharing system for reviews and ratings of its recognized accommodations. This information from the crowd is fundamental to both the marketing of any single lodging as well as to the quality of the service provided by the marketplace itself. This information also has a not negligible value in the fields of tourism economics and territory management. In fact, by considering data such as lodging prices, ratings, reviews, etc. in relation to their locations, it is possible to perform analyses on the territory attractiveness which provide valuable inputs for the implementation of proper territorial conservation as well as promotion policies [106]. The presented example focuses on the analysis of the spatial correlation between Airbnb© lodging average prices and ratings for the City of Venice. The first attribute is the average users' rating ranging from *0.0* to *5.0*, and the second is its average price per day in [€/day]. The input dataset included about

[3] https://www.airbnb.com

4000 point geometries describing lodgings locations and attributes. This data are not distributed with an open license and we were allowed by the provider to access it for research purpose only. Data refers to the Year 2015. The selected LISA was the bivariate Local Moran's I, computed for the lodging ratings as the base attribute and the prices. Due to the significant variation of points density within the study region, the spatial weights matrix was computed by using the KNN strategy, with K value set to 30. The resulting map is shown in Figure 5.2. In general, it can be observed that high values clusters (i.e. locations where high ratings are surrounded by high prices) are mostly concentrated in the city centre. Conversely, low values clusters are segregated in the peripheral areas. This is a typical situation in most of the historical cities where the centre embodies most of the tourist destinations. In Venice, this pattern is perfectly followed by the accommodation prices and ratings. Indeed, the LISA map provides valuable insights into the context of tourism management enabling to explore the location attractiveness or attractiveness disparities across urban landscapes.

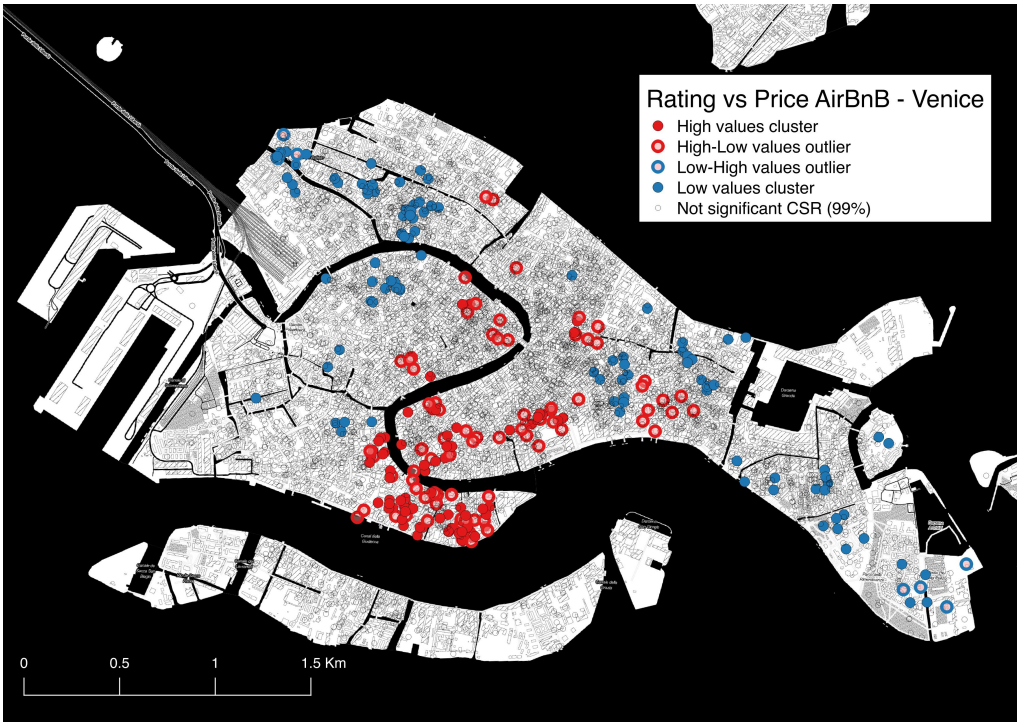


Figure 5.2: Resulting map from the computation of the bivariate Local Moran's I on the Airbnb© accommodation ratings and prices for the city of Venice (Italy) in 2015. *Basemap: ©OpenStreetMap contributors, ©Stamen Design.*

5.2.2 Case study: Soil consumption and incomes

In the second example, the bivariate Local Moran's I was employed using the Hotspot Analysis Plugin to investigate the spatial patterns of soil consumption in Italy and its interaction with a potentially linked macroeconomic variable, i.e., the average income per capita.

The phenomenon of soil consumption is a pressing concern within the research domains related to natural resources management. In fact, the soil provides humankind with most of the ecosystem services needed for its livelihood. At the same time, the limited awareness on the effects of its degradation, coupled with the long time necessary for the natural recovery of its functionalities and/or the prohibitive costs of its restoration, make the soil a fragile, limited, and non-renewable resource [32]. The principal causes of the soil consumption can be directly attributed to human-driven phenomena such as the urbanization, the demographic growth as well as the economic activities which take place along the territories. In Italy, the monitoring of soil consumption is performed by the Italian National Institute for Environmental Protection and Research (ISPRA)[4] with the support of the Copernicus program of the European Commission[5]. According to [96], in Italy the soil consumption affects about the 7 % of the whole territory (21.000 $[Km^2]$) with a rate of new soil consumed, which is defined as the changeover of a surface unit from natural to impervious soil per unit of time, close to 4 $[m^2/\text{s}]$ between years 2012 and 2015. Considering this time interval, the estimated economic impact in terms of loss of the main ecosystem ranges between 538.3 and 824.5 million [€]. The reported data highlight the relevance of this topic which is a geographically widespread phenomenon that needs to be properly described and understood taking advantage of geospatial technologies. For this reason, it represents a suitable case study for the application of LISA

[4] http://www.isprambiente.gov.it
[5] http://www.copernicus.eu

mapping.

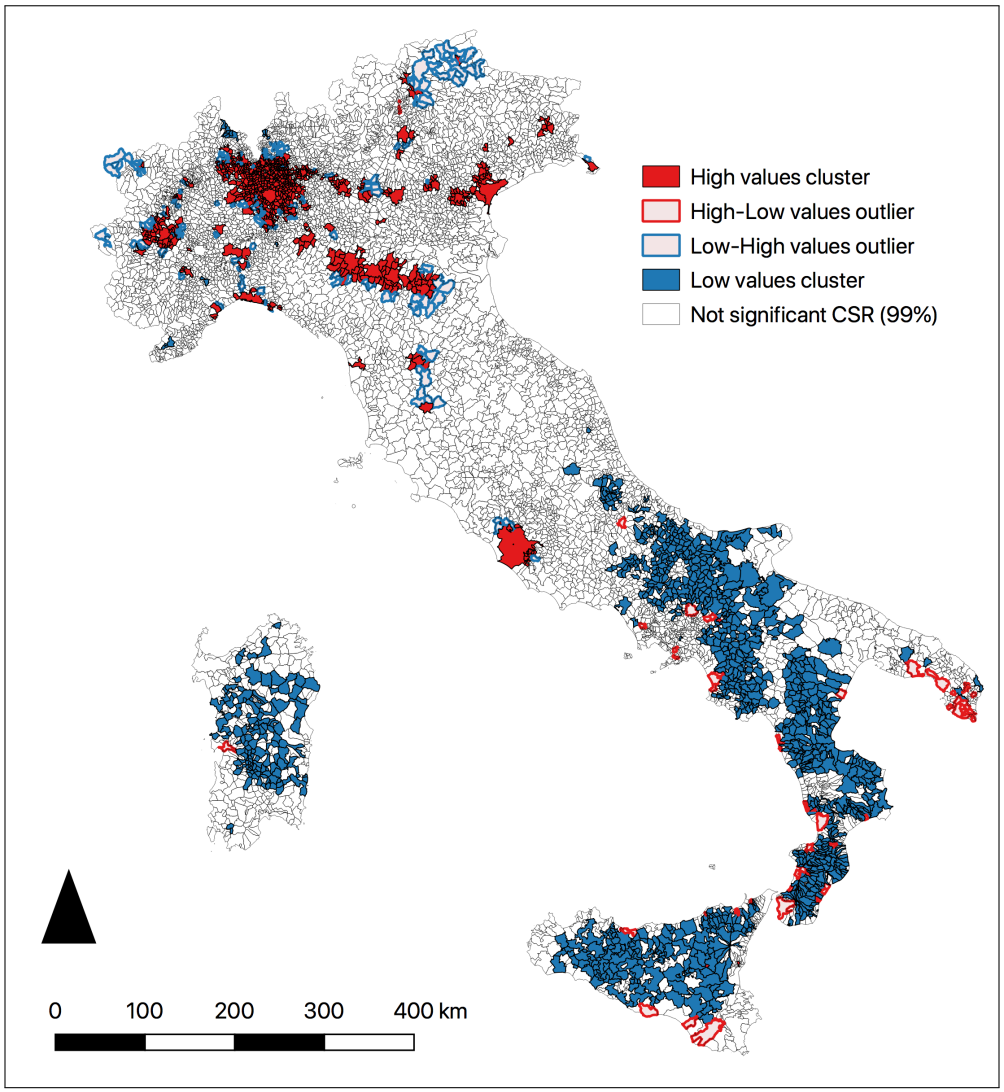


Figure 5.3: LISA map resulting from the computation of the bivariate Local Moran's I statistics for the total soil consumed and the average income per capita in the Year 2012.

The LISA mapping was performed to help in reply to the following questions: being the soil a resource, does its consumption generate in turn economic wealth? And if so, is this true everywhere along the national territory? [90]. Considered data were made available by Italian public institutions and released with a Creative Commons (CC-BY) open license. Since 2014, ISPRA distributes binary raster maps at 10 [m] pixel resolution depicting the soil consumption at the national scale. Whereas, the average income per capita [84]

was retrieved from the Italian Ministry of Economy and Finance Web portal[6]. A shapefile of Italian municipalities was enriched with the information of the average income per capita in the Year 2012 [€], and totally consumed soil until the same year as a percentage of each municipality area. The resulting bivariate Local Moran's I map computed from the two variables is included in Figure 5.3.

The bivariate cluster patterns depict a diffuse positive association between the two variables. High values cluster are present only the northern part of the country while low values clusters are confined in the Southern area. The well-known north-south disparity may drive this particular pattern in terms of incomes. Some exceptions can be observed. This is e.g., the case of Rome (Central Italy) where a high values cluster portrays a unique situation for this city concerning the neighbouring areas. It is interesting to observe how the outliers are distributed. In fact, almost all the detected LH outliers are located in Northern Italy while Hight-Low values outliers are present in Southern Italy only. Regarding Low-High values outliers, these show that in the north many cases can be found where low levels of soil consumption are spatially connected with high average incomes per capita. This is particularly true for Aosta and Bolzano provinces and the inland part of Tuscany Region. The observed patterns may be partially explained by the fact that mountain territories are generally less affected by soil consumption [25]. Nevertheless, the sharped negative association between soil consumption and income suggests that in these areas, the presence of a significant economic wealth may be not directly connected to the consumption of soil. A specular situation is detected in the South where many cases of high levels of soil consumption are spatially associated with low average incomes per capita, thus suggesting a more significant number of situations where the soil is consumed without a spatially associated generation of incomes. The obtained results produced a snapshot of the phenomenon of soil consumption

[6]http://www.mef.gov.it/en

in Italy, advising possible explanations of local critical issues, and highlighting evidence of disparities in the use of this resource along the territories.

Besides the specific case studies illustrated above, it has been spelt out the crucial role of both univariate and bivariate LISA in exploring characters of a large spatial dataset as well as in clearly presenting findings that are valuable for driving any further analysis. Exploratory results presented by means of maps can be used for load or dismiss credits to preliminary assumptions formulated on the data. With this in mind, both limits and constraints of map-based analysis using LISA are worth to be remarked. Detected patterns do not provide any power to explain the spatial processes which generated them. Patterns may be due to unobserved covariates as well as to not optimal settings of the exploratory experiments for the variable under investigation. Moreover, correlation does not imply any causation among the phenomena under study. Therefore, assumptions that are best supported by the LISA mapping require always to be validated through confirmatory procedures such as regressions, variance analyses, etc.

Chapter 6

Multivariate Spatial Association Analysis

In this chapter, an application of the multivariate Local Geary's c (Chapter 3) is presented. This technique is nested into an experimental procedure enriching the outcomes of the multivariate LISA by means of additional data analysis techniques as well as descriptive statistics. The proposed procedure provides a reproducible workflow enabling multivariate clusters and outliers classification and mapping (Section 6.1). This early application report is intended to fulfil the primary research objective of extending traditional LISA analysis and mapping to the multivariate context. As a case study, the multivariate analysis of some social vulnerability indices for the City of Melbourne (Victoria, Australia) is presented and discussed in Section 6.2.

Most of the discussion of local spatial association has been situated in a univariate context. The treatment of spatial association in a multivariate setting has focused mainly on global statistics, specifically the Moran's I [111]. Recently, an extension of traditional LISA techniques (Equation 3.13) has been proposed by [8]. This is derived from the univariate Local Geary's c statistic

(3.8), introduced by [4], which allows for identifying multivariate local spatial clusters and outliers. However, the multivariate Local Geary's c, like its univariate version, does not provide any possibility to classify cluster and outlier, i.e. into high or low values clusters, etc. such as for the others LISA. This is a limiting factor in the exploration and mapping of multivariate spatial patterns.

In the next sections, an experimental procedure is proposed to enable multivariate spatial clusters and outliers classification. The procedure is based on the multivariate Local Geary's c computation (Chapter 3). The obtained clusters and outliers map is post-processed to extract auxiliary indicators allowing both to get insight into the multivariate clusters intensity as well as into the types of cluster and outliers, such as high, low, etc. These tasks are plugged into a pipeline, implemented through custom Python scripts, that provides as output multivariate classified cluster and outlier maps which are comparable to the output of the univariate LISA mapping. The procedure is tested on the City of Melbourne to investigate the multivariate spatial association of some social vulnerability indices. Results are validated through the use of independent data.

6.1 Procedure outline

Data preparation

The designed procedure for multivariate spatial association mapping requires as input a geospatial layer depicting locations within a region of interest. At each of the n locations is assigned a tuple of numerical attributes representing observations of the k variables for which the multivariate spatial association patterns need to be investigated. This layer can be e.g. a shapefile including n geometries and k variables into separate fields of the attribute table. High

variability in the ranges of analysis variables is likely to be encounter in a multivariate setting. Therefore, a Z-score scaling standardization is applied to each variable observations k_i (Equation 3.6) to allow for comparisons.

Multivariate LISA computations

The standardized attributes z_{k_i} are used to compute the multivariate Local Geary's c at each location in the dataset (c_{k_i}) (Equation 3.13). A spatial weight matrix for modelling the neighbour relationship among the n locations in the dataset needs to be defined by the analyst and included in the computation.

Inference on the CSR hypothesis is provided through conditional random permutation tests by assigning a pseudo p-value p_i (Equation 3.3), eventually corrected by means of FDR procedures, such as the Benjamini-Hochberg procedure (Equation 3.4), to each location in the dataset. Significant pseudo p-values are used to map interesting locations thus focusing on candidate spatial clusters or outliers.

Spatial association type definition

The definition of the spatial association type, i.e. positive or negative, for each interesting location is achieved by comparing the value of the c_{k_i} computed from observations, with its expected value that can be either the observed sample mean or the mean of its empirical reference distribution derived by the conditional permutation tests. c_{k_i} values higher than the expected value depict negative spatial association while lower values positive association. By coupling this information with pseudo p-values, a map of the multivariate spatial clusters and outliers for the considered variables is obtained. The above steps encompass the general procedure for the Local Geary's c mapping whereas the following ones are introduced to tackle the description of the spatial association type for

the detected multivariate clusters and outliers. This LISA is key to the proposed procedure. Indeed, it allows for testing on CSR in a multivariate setting, thus focusing the additional processing, perhaps not directly embedding spatial association measures, only on those locations for which spatial association is likely to be present.

Test on PCA projections

A parallel test for spatial association can be implemented by exploiting the output of the multivariate Local Geary's c mapping. This consists of projecting the attribute tuples of each location on a reduced attribute space by mean of PCA [70]. By identifying on this new space the position of each interesting location i, i.e. clusters and outliers, and the position of its neighbours j, it is possible to retrieve a simple measure of dispersion that provides insight on the cluster intensity. The PCA is a well-established EDA technique that can be used for dimensionality reduction in multivariate analyses. Basically, the PCA maps a number of possibly correlated variables into a smaller (or equal) number of linearly uncorrelated variables by means of orthogonal transformations of the original variable axes to new orthogonal axes called Principal Components PC. These new axes coincide with directions of maximum variation for the original observations such as that the PC1 corresponds to the direction of maximum variation, the PC2 to the second direction of maximum variation, and so on. Therefore, PCA is a decomposition of the $n \times k$ matrix $\mathbf{X}$ containing the n standardized observations of the k original variables, such as in Equation 6.1 [35].

$$\mathbf{T} = \mathbf{XP} \leftrightarrow \mathbf{X} = \mathbf{TP}^T \tag{6.1}$$

Where $\mathbf{T}$ is the $n \times q$ matrix containing the projections of $\mathbf{X}$ onto the new space defined by the q considered PC, with $q \leq k$. The transformation $\mathbf{X} \rightarrow \mathbf{T}$ is computed by means of an orthogonal projection matrix $\mathbf{P}$ such as $\mathbf{P}\mathbf{P}^T$ is equal to the identity matrix $\mathbf{I}$. Matrix $\mathbf{P}$ is often referred as *loading matrix* and $\mathbf{T}$ as *score matrix*. $\mathbf{P}$ is computed so that its columns are the directions of maximum variance in the data, with the first column (or PC1) representing the direction of maximum variance, the second column (PC2) the direction of the next largest variance, and so on. These directions correspond to the eigenvectors of the correlation matrix $\mathbf{S}$ computed from the standardized observations matrix $\mathbf{X}$ as in Equation 6.2.

$$\mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{n-1} \tag{6.2}$$

Ordering the eigenvectors of $\mathbf{S}$ so that the correspondent eigenvalues are in descending amplitude order, Equation 6.3 is verified.
Λ corresponds to $\text{diag}(\lambda_{S_1}, \lambda_{S_2}, \cdots, \lambda_{S_k})$ winch is the diagonal matrix of the ordered eigenvalues of $\mathbf{S}$.

$$\mathbf{P}^T\Lambda\mathbf{P} = \mathbf{S} \tag{6.3}$$

From Equation 6.3, it follows that each PC corresponds to the direction of one eigenvector and is a linear combination of the original variables. The percentage of the total variance explained by a generic principal component PC_q is equivalent to its corresponding eigenvalue λ_{S_q} divided by the sum $\sum_{v=1}^{k} \lambda_v$ of all eigenvalues of $\mathbf{S}$.

In the proposed procedure, n observations for each k variable are considered. The PCA is leveraged for reducing k and providing a compact view of the attribute space preserving the variability of the observations. Suppose to consider $q = 2$, thus projecting the $n \times k$ standardized observations matrix $\mathbf{X}$ onto a new bi-dimensional space defined by the direction the first two principal components PC1 and PC2. We call $\mathbf{P}_2$ the matrix containing the first two columns of the full loading matrix $\mathbf{P}$, and $\mathbf{T}_2$ the corresponding scores matrix. Hence Λ_2 being equal to $\text{diag}(\lambda_{S_1}, \lambda_{S_2})$ with $\mathbf{S}$ denoting the correlation matrix from $\mathbf{X}$. Therefore, $\mathbf{T}_2$ contains the coordinates of the original observations mapped onto the plane defined by PC1 and PC2. The percentage of the total variance explained by $\mathbf{T}_2$ is given by Equation 6.4.

$$\frac{\sum_{v=1}^{q} \lambda_{S_v}}{\sum_{v=1}^{k} \lambda_{S_v}} \tag{6.4}$$

Where $\mathbf{S}$ is the $k \times k$ correlation matrix from the $n \times k$ standardized observations matrix $\mathbf{X}$, and λ_S are the eigenvalues of $\mathbf{S}$. Providing that the percentage of the total variance explained by $\mathbf{T}_2$ is large enough to meet the (arbitrary) analyst's needs, a compact and meaningful representation of the original k variables is obtained by means of the bi-dimensional space PC1 - PC2. In many cases, $\mathbf{X}$ can be decomposed using a small number q of PC, with $q << k$, while still explaining most of the variance in the data [35].

It is now possible to isolate the projections of both each interesting location i, outlined by the multivariate Local Geary's c, and its neighbours j on the PC1 - PC2 plane. Neighbours are defined according to the spatial weights matrix adopted in the multivariate Local Geary's c computation. For each interesting location i, a specific $r \times q$ subset $\mathbf{T}_{2,i,j}$ of the $\mathbf{T}_2$ elements is considered. These

elements consist of projected coordinates $PC1_i$ and $PC2_i$ of the original observations at location i and at its neighbours j that resulted in a spatial clusters or outlier from the multivariate LISA mapping. To explore the spatial association characterizing these significant locations in the PCA projected attribute space, a simple measure of local dispersion D_i is suggested as follows. Denoting $t_{PC1_i,j}$ and $t_{PC2_i,j}$ respectively the first and the second column elements of the subset $\mathbf{T}_{2,i,j}$, the coordinates of the centroid (or mass centre) V_i on the PC1 - PC2 plane for the points of the subset can be computed as in Equation 6.5.

$$V_i = \begin{cases} V_{i,PC1} = \frac{\sum_{j=1}^{r} t_{PC1_i,j}}{r}, & V_i \text{ coordinate on PC1} \\ V_{i,PC2} = \frac{\sum_{j=1}^{r} t_{PC2_i,j}}{r}, & V_i \text{ coordinate on PC2} \end{cases} \tag{6.5}$$

By computing the average distance from V_i to each point $(t_{PC1_i,j}; t_{PC2_i,j})$ a local measure of dispersion (or cluster intensity) D_i is obtained (Equation 6.6).

$$D_i = \frac{\sum_{j=1}^{r} \sqrt{(V_{i,PC1} - t_{PC1_i,j})^2 + (V_{i,PC2} - t_{PC2_i,j})^2}}{r} \tag{6.6}$$

According to the definition of the D_i, the closer the value to 0 the higher the cluster intensity. The D_i computed at each interesting location can be connected to the type of spatial association, i.e. positive or negative, characterising the considered locations. Accroding to this logic, low values for D_i are expected for clusters whereas higher values for outliers.

The $\mathbf{T}_{2,i,j}$ subset from the PCA is linearly uncorrelated from the original variable tuples at locations i and j used to compute the multivariate Local Geary's

c. Hence D_i values can be used to validate the cluster and outlier obtained from the multivariate LISA. A simple validation test can be performed by observing if the behaviour of the D_i is in agreement with the reference multivariate cluster and outlier map. Therefore, the D_i test can be considered an internal validation, i.e. directly derived from observations, of the spatial association type defined by means of the multivariate Local Geary's c. Moreover, the D_i values provide an exploratory and mappable metric to describe the intensity of detected clusters and outliers, which can be displayed, e.g. on the LISA map using a composite visualisation style. The use of only two principal components eases the visual exploration of results from the PCA projections test through bi-dimensional scatter plots. An example of the above is included in the next section and additional considerations on the D_i are illustrated in Appendix B. Nevertheless, the test can be run using more than two principal components in the case of an unsatisfactory percentage of the total variance explained by PC1 and PC2.

Multivariate clusters and outliers classification

Due to their mathematical formulation, neither the multivariate Local Geary's c nor the suggested local measure of dispersions D_i allow to classify the type of clusters and outliers such as other LISA e.g. the Local Moran's I. This operation can be performed by means of linking and brushing operation with univariate LISA maps as suggested by [8]. However, the purpose of the presented procedure is to outline an automatic classification strategy for multivariate clusters and outliers by exploiting descriptive indicators computed from the input data. The designed strategy is based on the comparisons of centrality measures from the sample distributions of the considered variables. As for the case of PCA, these comparisons are constrained to the significant locations detected by means of the multivariate Local Geary's c. The classification logic is inherited from the Local Moran's I (Equation 3.7).

Generally speaking, a cluster is detected where observations at a focal location i are similar to those in the neighbouring locations j. A cluster is labelled as HH (high values) when this observations subset includes higher values than the expectation of the correspondent variables in the study region. A cluster is labelled as LL (low values) when the subset shows lower values than the expectation. An outlier is instead detected where observations at a focal location i are dissimilar to those in the neighbour locations j. An outlier is labelled as HL (high values surrounded by low values) where the observations at a focal location i are higher than the expectation of the correspondent variables in the neighbour locations j. An outlier is labelled as LH (low values surrounded by high values) where the observations at a focal location i are lower than the expectation of the correspondent variables in the neighbour locations j. The definition of high and low values has a local validity and it is connected to the ranges of the considered variables thus may differ from case to case. Indeed, no threshold values are specified in the above description, and all the comparison are referred to local expectation, e.g. the sample mean for the analysis variables.

The empirical measure Mm_c enabling the classification of multivariate clusters is proposed as follows. $\mathbf{X}$ is the $n \times k$ matrix containing the standardized observations of the k original variables at each of the n locations. For a location i that resulted in a cluster from the multivariate Local Geary's c, a $r \times k$ subset $\mathbf{X}_{i,j}$ containing the observations at location i and at its neighbours j can be extracted. The idea is to compare the mean of the vector of the column-wise medians of $\mathbf{X}_{i,j}$ ($\mu_{\mathbf{X}_{i,j}^M}$) with the mean of the vector of the column-wise medians of $\mathbf{X}$ ($\mu_{\mathbf{X}^M}$). This implies comparing the local mean of the medians for the considered variables at a cluster location with the ones of the whole study region. The median is preferred to the mean for performing the column-wise aggregation because to account for possible skewed distributions of the analysis

variable. This reduces the bias in estimating the central tendency values for the comparison, due to the influence of possible outliers [109]. The proposed measure $Mm_{c,i}$ consists of a simple difference between $\mu_{\mathbf{X}^M_{i,j}}$ and $\mu_{\mathbf{X}^M}$ at each cluster location i (Equation 6.7). Positive values of the $Mm_{c,i}$ depicts a higher local mean of the medians than the one of the whole study region, hence a possible HH cluster. Negative values of the $Mm_{c,i}$ depicts a lower local mean of the medians than the one of the whole study region, hence a possible LL cluster. The absolute value of the $Mm_{c,i}$ provides an indicator for performing inter-comparisons among clusters. Indeed, the higher the absolute value of the $Mm_{c,i}$, the stronger the HH or LL relationship with respect to the average conditions for the study area.

$$Mm_{c,i} = \mu_{\mathbf{X}^M_{i,j}} - \mu_{\mathbf{X}^M}, \rightarrow \begin{cases} Mm_{c,i} > 0, & \text{HH} \\ Mm_{c,i} \leq 0, & \text{LL} \\ |Mm_{c,i}|, & \text{HH or LL intensity indicator} \end{cases} \tag{6.7}$$

The empirical measure Mm_o enabling the classification of multivariate outliers is proposed as follows (Equation 6.8). $\mathbf{X}$ is the $n \times k$ matrix containing the standardized observations of the k original variables at each of the n locations. For a location i that resulted in a outlier from the multivariate Local Geary's c, two subsets $\mathbf{X}_i$ and $\mathbf{X}_j$ containing respectively the $1 \times k$ observations at location i and the $r-1 \times k$ observations at its neighbours j can be extracted. In this second case, the idea is to compare the mean of the vector $\mathbf{X}_i$ ($\mu_{\mathbf{X}_i}$) with the mean of the vector of the column-wise medians of $\mathbf{X}_j$ ($\mu_{\mathbf{X}^M_j}$). This implies comparing the mean of the observations at an outlier location with the mean of the medians of its neighbour locations only. The median is again suggested for the same reasons explained before. The proposed measure $Mm_{o,i}$ consists of a simple difference between $\mu_{\mathbf{X}_i}$ and $\mu_{\mathbf{X}^M_j}$ at each outlier location i (Equation 6.8).

Positive values of the $Mm_{o,i}$ depict a higher local mean than the mean of the medians from the neighbour locations, hence a possible HL outlier. Negative values of the $Mm_{o,i}$ depict a lower local mean than the mean of the medians from the neighbour locations, hence a possible LH outlier. The absolute value of the $Mm_{o,i}$ provides an indicator for performing inter-comparisons among outlier. Indeed, the higher the absolute value of the $Mm_{o,i}$, the stronger the HL or LH relationship detected.

$$Mm_{o,i} = \mu_{\mathbf{X}_i} - \mu_{\mathbf{X}_j^M}, \rightarrow \begin{cases} Mm_{o,i} > 0, & \text{HL} \\ Mm_{o,i} \leq 0, & \text{LH} \\ |Mm_{o,i}|, & \text{HL or LH intensity indicator} \end{cases} \tag{6.8}$$

Results form the Mm_c, and the Mm_o computations can be used to enrich clusters and outliers map obtained from the multivariate Local Geary's *c*. With this additional information, clusters and outliers can be classified thus producing a multivariate LISA map comparable to that obtained, e.g. from the Local Moran's *I* in the univariate analysis. Additional measures such as the absolute values of $Mm_{c,i}$ and $Mm_{o,i}$ might be also included and displayed on the multivariate map favouring the exploration of the spatial association features of the k considered spatial variables. The full analysis workflow is summarised in Figure 6.1. Despite the relatively large number of steps required by the proposed procedure, the definition of an automatic strategy for enriching multivariate spatial association maps is achieved. Being the proposed procedure highly experimental and not exhaustively tested, results have to be intended as an exploratory tool rather than in a rigorous statistical sense. The same applies to the LISA technique on which the whole procedure is based, i.e. the multivariate Local Geary's *c*.

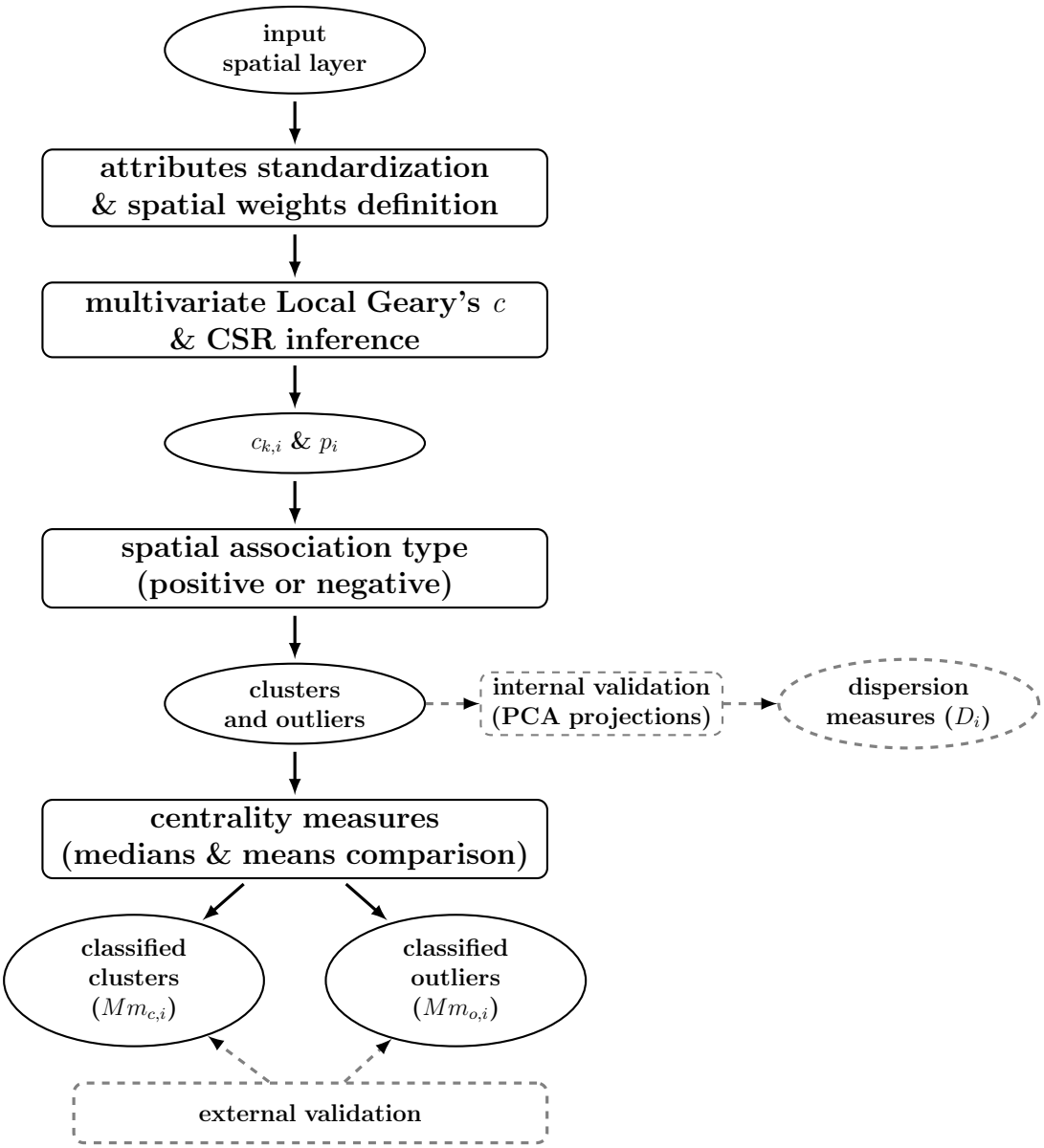


Figure 6.1: Flow chart of the proposed procedure for multivariate clusters and outliers classification and mapping. Inputs and outputs are marked into ellipses, full line boxes include the principal tasks of the procedure and dashed boxes represent suggested additional steps.

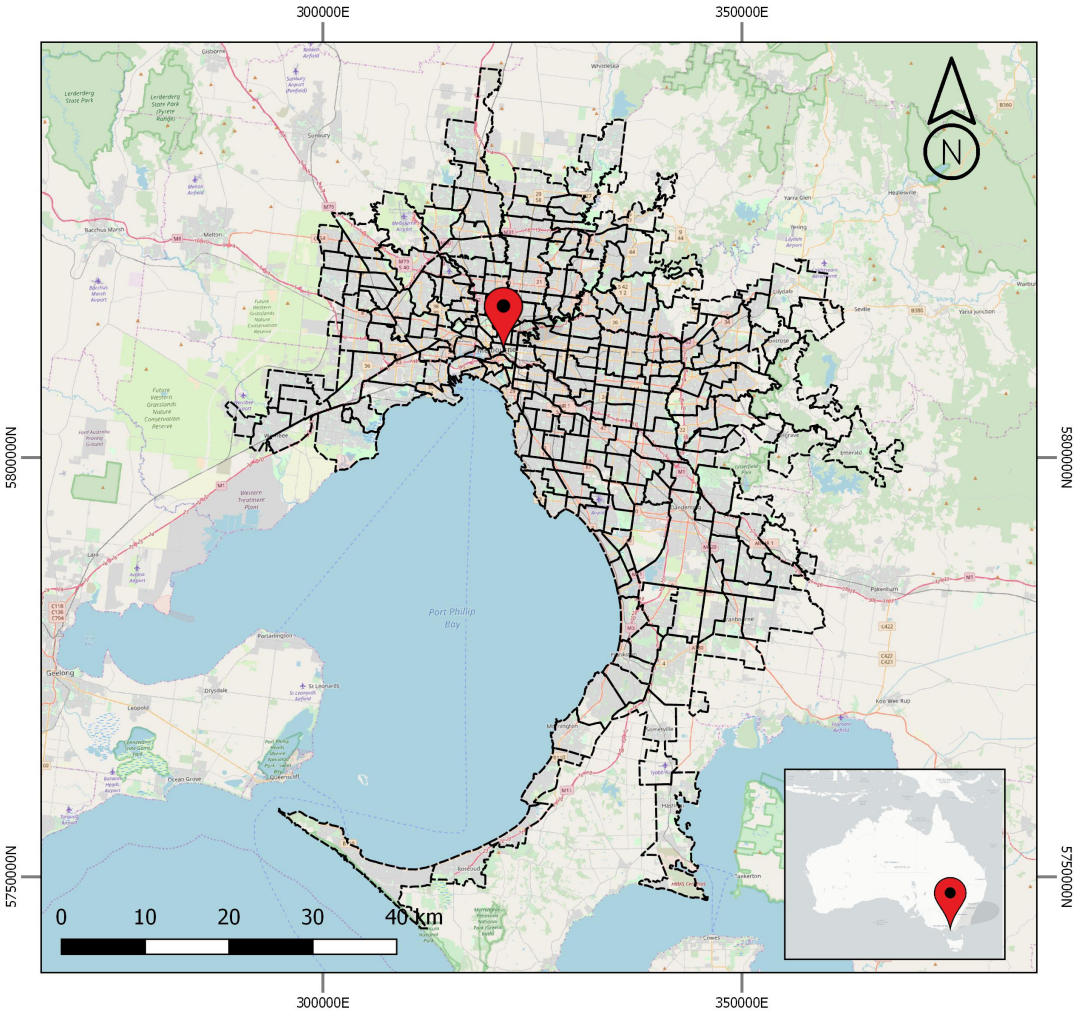
In views of the above considerations, either a visual or numerical validation of the results is strongly suggested. A visual validation can be achieved, e.g. by observing choropleth maps of the original observations where also complex spatial patterns might be easily detected by eye. Therefore, a careful visual inspection and interpretation of the detected pattern should always accompany the analysis [110]. Based on data availability, the procedure might be repro-

duced by involving independent variables which are semantically or numerically linked to the one used for the analysis. This might provide more robust confirmations for the analysis results. A sample validation is presented in Section 6.2. However, no validation procedures based on independent or reference data with a general validity are identified within this work.

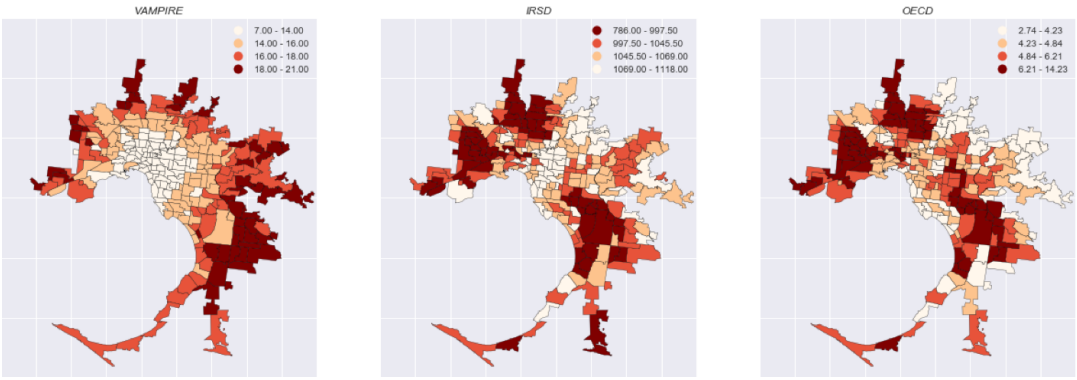
6.2 Case study: Social vulnerability in the City of Melbourne

In the presented case study, the procedure outlined in the previus section is tested for analysing and mapping of multivariate spatial association patterns of some social vulnerability indices available for the City of Melbourne.

Over the past two decades, Melbourne has been in the midst of a third great demographic change that rivals the Australian Gold Rush (mid to late 19th Century) and the post Second War boom. The more the city sprawls, the greater the risk it will become an unsustainable city divided by disadvantage and inequity [57]. The purpose is to automatically identify interesting or critical spatial patterns for the social vulnerability which might be further investigated to deliver valuable insights into city planning and decision making practices. To accomplish that, we considered three numerical variables representing indices of social vulnerability such as the Vulnerability Assessment for Mortgage, Petrol and Inflation Risks and Expenditure (VAMPIRE), the Index of Relative Socio-economic Disadvantage (IRSD), and the Total Unemployment Rate. Additional information for the considered indices is available in Table 6.1. Data refer to the Year 2011. The indices have been selected to account for some of the most relevant factors contributing to the social vulnerability in an urban context. Among others, employment conditions, housing, health, education, and purchase power [100]. The selection of the reference year is justified by the



(a)



(b)

Figure 6.2: (**a**) reference map of the City of Melbourne with SA2 boundaries (CRS: WGS84-UTM55S). The red marker identifies the central business district. (**b**) spatial distribution of the considered social vulnerability indices in the Year 2011. The visualization style is based on the fourth quantile break. *Basemap: © OpenStreetMap contributors.*

Acronym	Description *(Reference Year, Source)*
VAMPIRE	Vulnerability Assessment for Mortgage, Petrol and Inflation Risks and Expenditure. The average VAMPIRE score by definition is 15 out of 30. A low score indicates good performances in terms of social vulnerability. *(2011, http://www.vampire.org.nz/vampire)*
IRSD	The Index of Relative Socio-economic Disadvantage is a general socio-economic index that summarises a range of information about the economic and social conditions of people and households. A low score indicates a relatively greater disadvantage. In the analysis, the inverse of the IRSD scores is considered to disambiguate its interpretation with respect to the other two indices. *(2011, http://www.abs.gov.au)*
OECD	The total unemployment rate expressed as a percentage of the total labour force, where the latter consists of the unemployed plus those in paid or self-employment. Intuitively, a low score indicates good performances in terms of social vulnerability. Data is derived from the Organization for Economic Co-operation and Development (OECD) records. *(2011, https://data.oecd.org/unemp/unemployment-rate.htm)*

Table 6.1: Social vulnerability indices used in the multivariate analysis.

availability of consolidated information, such as census data and social indices, which was not yet dispatched by more recent surveys at the beginning time of this study. The indices are provided by census parcels (polygons) at the Statistical Area Level 2 (SA2) according to the Australian Statistical Geography Standard (ASGS). The study region includes the whole SA2 parcels (260) of the Melbourne metropolitan area, known as the Greater Melbourne (Figure 6.2a). The VAMPIRE and the IRSD are provided by the Australian Bureau of Statistics (ABS)[1] while the Total Unemployment Rate is distributed by the Organization for Economic Co-operation and Development (OECD)[2].

According to the mathematical definition of each index, a first manipulation of the data is performed. Namely, the scales of the variables have to be adjusted to meet the analysis requirements. In this case, hight values of both the

[1]http://www.abs.gov.au
[2]http://www.oecd.org

VAMPIRE and the OECD indicate lousy performance in terms of social vulnerability for a SA2 parcel whereas the lower values, the better performance. Conversely, low values for the IRSD means higher social vulnerability performances whereas hight values lower performance. Therefore, the inverse of the IRSD observations is considered in the computations. The spatial distribution of the considered indices is reported in Figure 6.2b where the colour ramp for the IRSD is inverted to graphically account to what stated above.

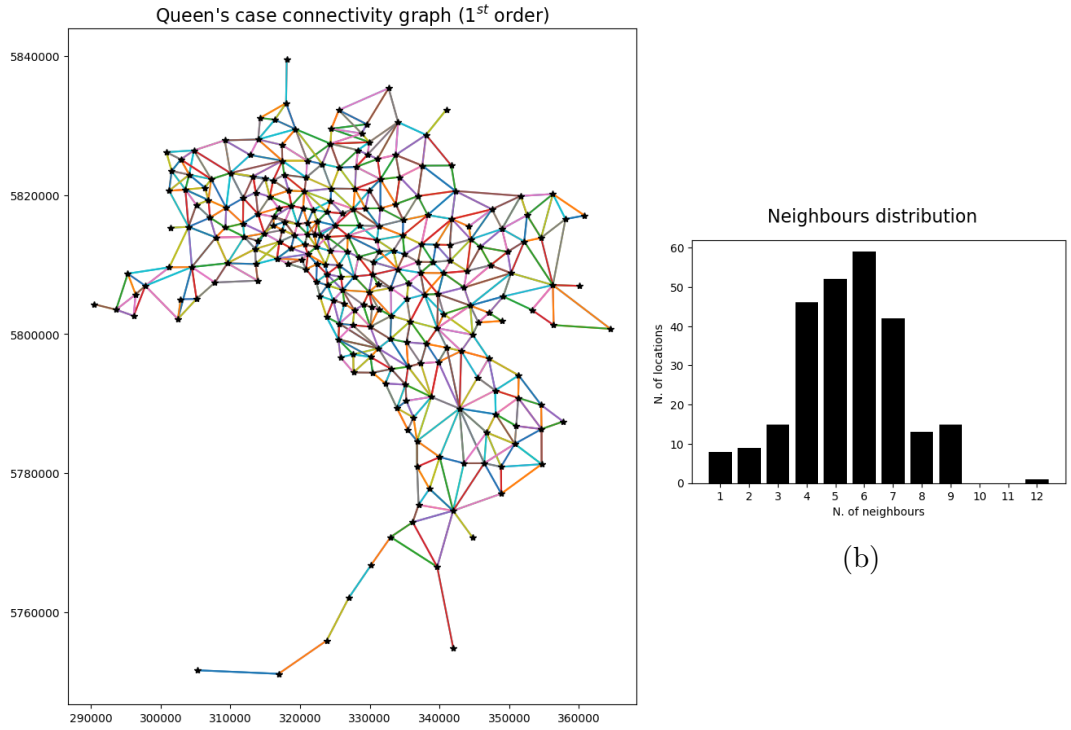
After a visual inspection of the spatial distribution of the indices, their global spatial association is assessed by means of the univariate and the bivariate Global Moran's I [75] (Table 6.2). The spatial weights matrix selected for this case study is based on the first order queen's case contiguity, i.e. edge and corner contiguity (Figure 6.3a). This guarantees a generally even distribution of neighbours (Figure 6.3b) excluding the presence of any island, i.e. disconnected locations. A row-standardized version of the obtained spatial weights matrix is used in the computations.

A global positive spatial association is detected according to the univariate Global Moran's I (diagonal values of Table 6.2), that is stronger for the VAMPIRE. The positive association also emerges from the bivariate Global Moran's I (extra-diagonal values of Table 6.2), especially between OECD and IRSD. This evidence can be partially linked to the fact that the IRSD index includes among its base indicators also the Total Unemployment (OECD).

The mapping of local multivariate spatial association patterns is performed by means of the Local Geary's c. The significance level α is arbitrary set equal to 0.001 and inference on CSR is performed using 9999 conditional permutations. Pseudo p-values are corrected using the Benjamini-Hochberg FDR procedure. Spatial association type is defined through comparisons between the computed values of the $c_{k,i}$ and its sample mean. Clusters and outliers are thus detected

	VAMPIRE	IRSD	EOCD
VAMPIRE	0.811	0.186	0.025
IRSD	0.175	0.538	0.456
EOCD	0.017	0.458	0.476

Table 6.2: Global Moran's I matrix computed from the considered indices. Diagonal values correspond to the univariate Global Moran's I, and extra-diagonal values result from pairwise comparisons between indices by means of the bivariate Global Moran's I.



(a) (b)

Figure 6.3: (**a**) connectivity graph for the Greater Melbourne based on the first order queen's case contiguity rule. The centroids of each SA2 parcel is used to represent the corresponding polygon. (**b**) histogram of the neighbours distribution.

and mapped. In total, 27 significant locations are spotted of which 20 resulted in being clusters and 7 outliers. Spatial association patterns for social vulnerability indices in the Greater Melbourne are summarised in Figure 6.4.

A parallel assessment of the multivariate Local Geary's c outcomes is performed using PCA projections and the D_i measure of dispersion, proposed in Section 6.1. The first two PC are considered which explain about 95% of the total

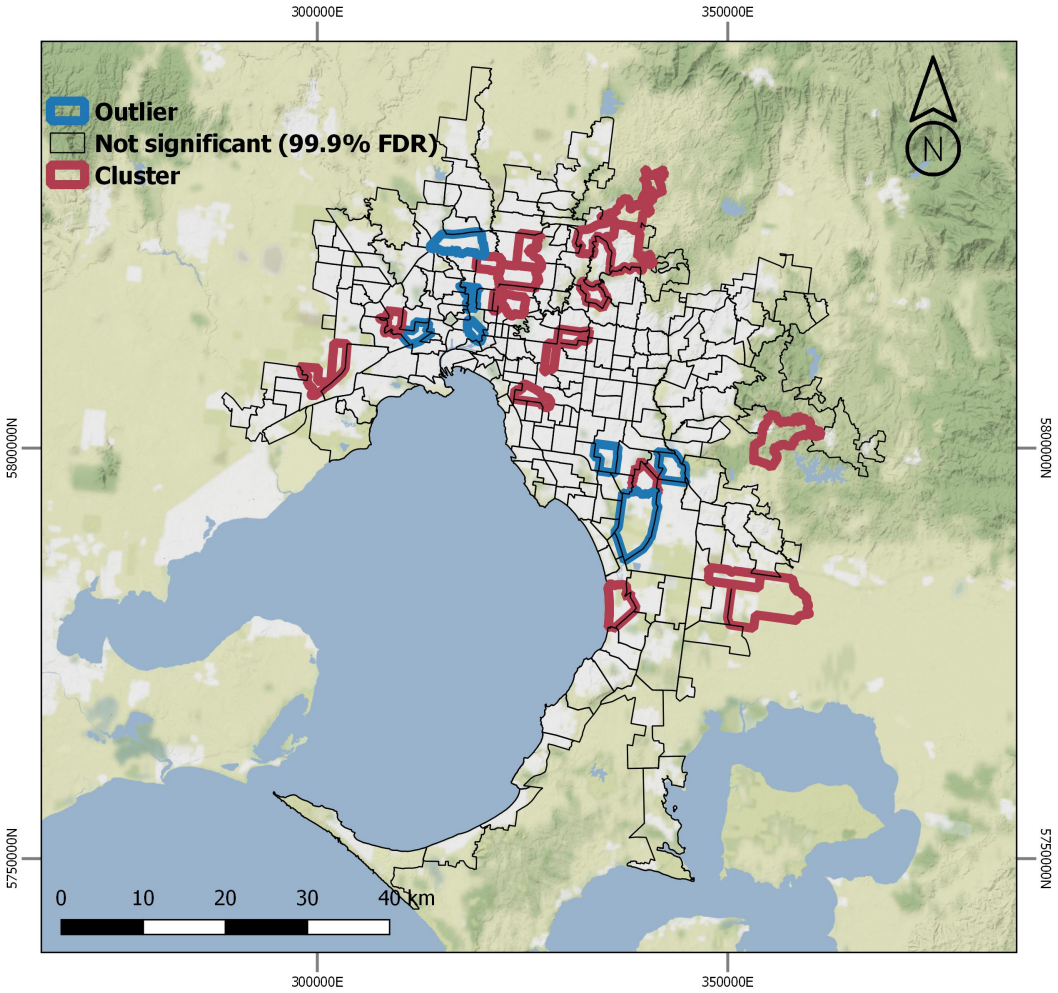


Figure 6.4: Resulting clusters and outliers map from the computation of the multivariate Local Geary's c for the three selected social vulnerability indices. *Basemap: ©OpenStreetMap contributors, ©Stamen Design.*

variance for the three variables considered. The resulting loading matrix $\mathbf{P}$ is reported in Table 6.3.

The D_i is computed using the same spatial weights matrix described before to model the neighbouring relationships among locations. Results are shown in Figure 6.5 and Figure 6.6. In Figure 6.5 the D_i measures are mapped over the correspondent significant locations. This metric allows for assessing the local degree of clustering in terms of dispersions from a local centre of mass in the PC reduced attribute space. Figure 6.6 provides a graphical insight on the D_i computations (Figure 6.6a, 6.6b) and its general agreement with the outcomes of the Local Geary's c. As expected, clusters show generally lower

	PC1	PC2
VAMPIRE	0.0847	-0.9851
IRSD	0.7091	-0.0458
EOCD	0.6999	0.1656

Table 6.3: Loading matrix **P** containing the two principal component loads for the three considered indices.

D_i than outliers. The only disagreement between the expected and the actual behaviours of the D_i is found in the middle of its distributions where an outlier shows a lower D_i than a cluster (Figure 6.6c). Further considerations on these issues as well as on the possible uses of the D_i as a measure of multivariate spatial association are discussed at the end of this section and in Appendix B.

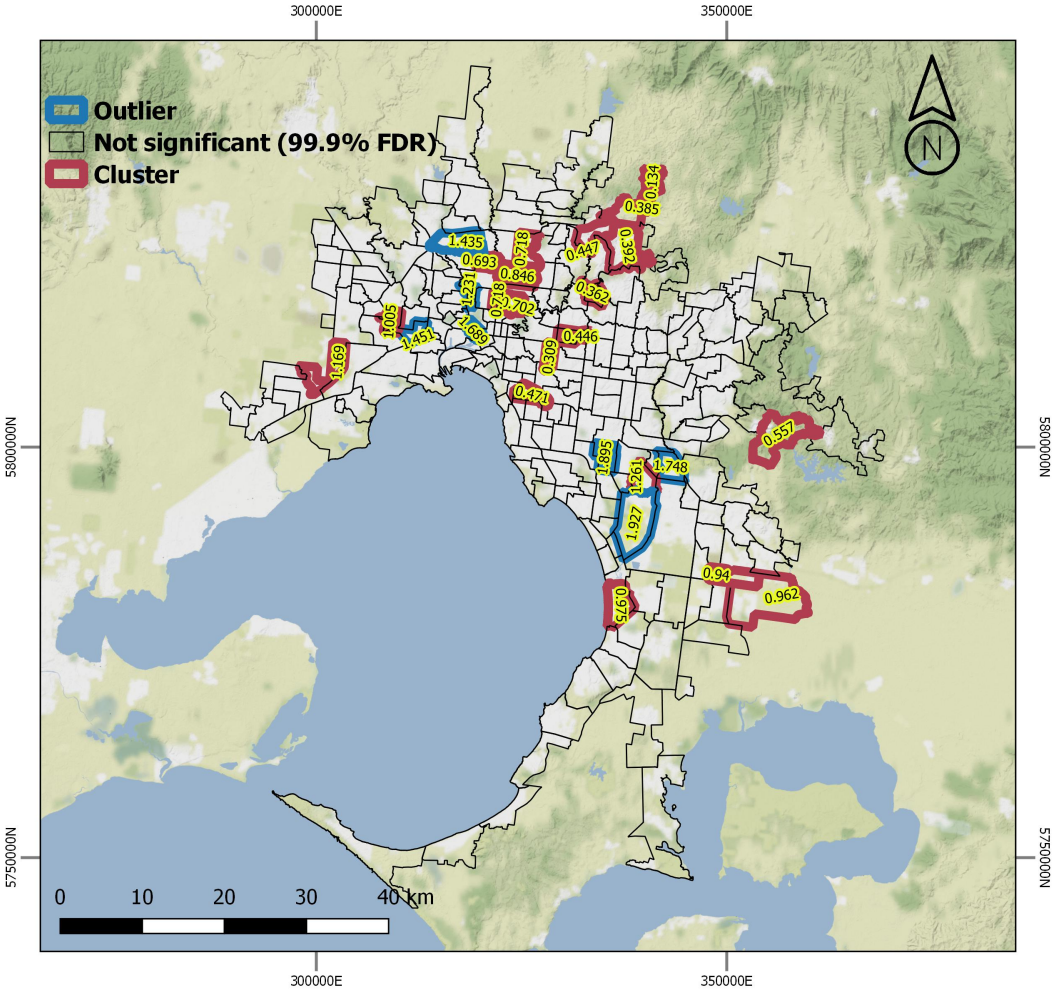
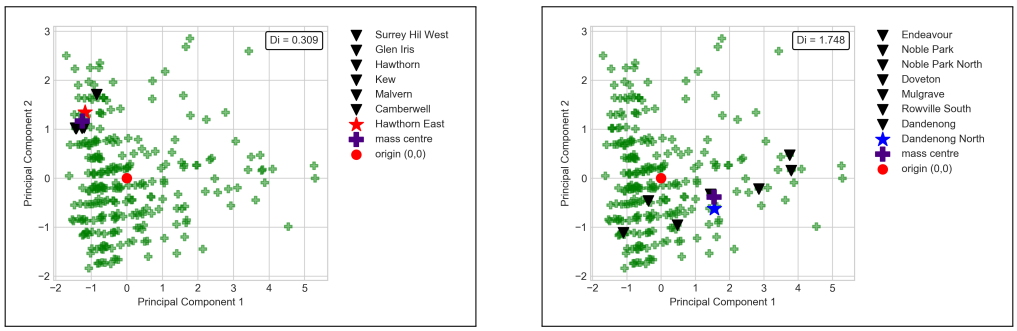
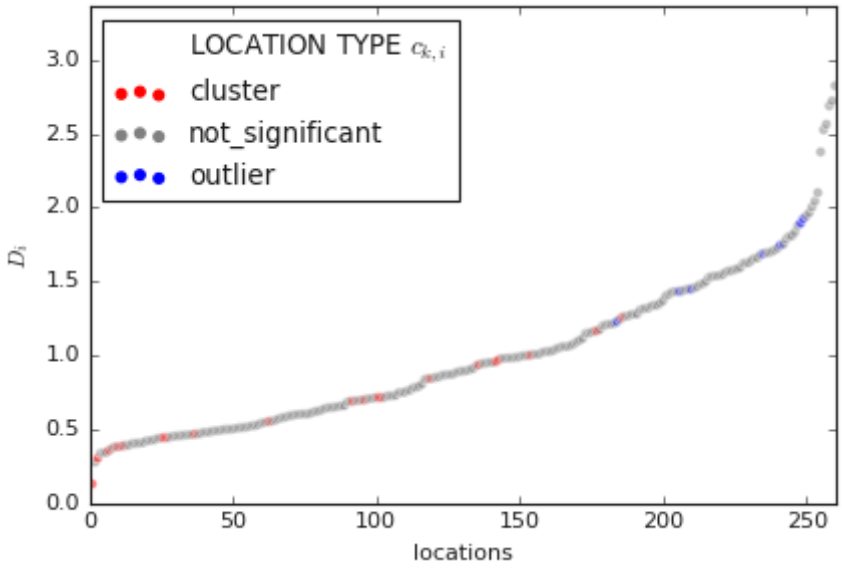


Figure 6.5: Resulting clusters and outliers map from the computation of the multivariate Local Geary's c enriched with the computed D_i measures. *Basemap: ©OpenStreetMap contributors, ©Stamen Design.*



(a) (b)



(c)

Figure 6.6: (**a**), (**b**) scatter plots of projected indices onto the PC plane (green crosses). The D_i measures for (**a**) a cluster location (red star) and (**b**) an outlier location (blue star) are highlighted together with their neighbours (black triangles) based on which the mass center is computed. (**c**) increasing ordered D_i measures for all the locations. Resulting clusters and outliers from the Local Geary's c are highlighted respectively with red and blue dots.

To complete the multivariate spatial association experiment, clusters and outliers classification is performed employing the Mm measures (i.e. means and medians comparison), proposed in Section 6.1. Results of the multivariate clusters and outliers classification are reported in Figure 6.7. This map provides insight on the type of clusters and outliers following a similar classification logic such as the one adopted by the Local Moran's I. Each cluster is labelled as high or low values according to the difference between the mean of the medians characterising observations in the cluster sub-region and the mean of the medi-

ans from the whole observations in the region (Mm_c). Each outlier is instead labelled as high values surrounded by low values, and vice versa, according to the difference between the mean characterising observations at the outlier location and the mean of the medians from the observations at its neighbours (Mm_o). The absolute value of this difference can be adopted as a measure of intensity for clusters and outliers in terms of dissimilarity between central tendencies of local observations and the one characterising either the study region or the neighbours, depending on the type of significant locations considered.

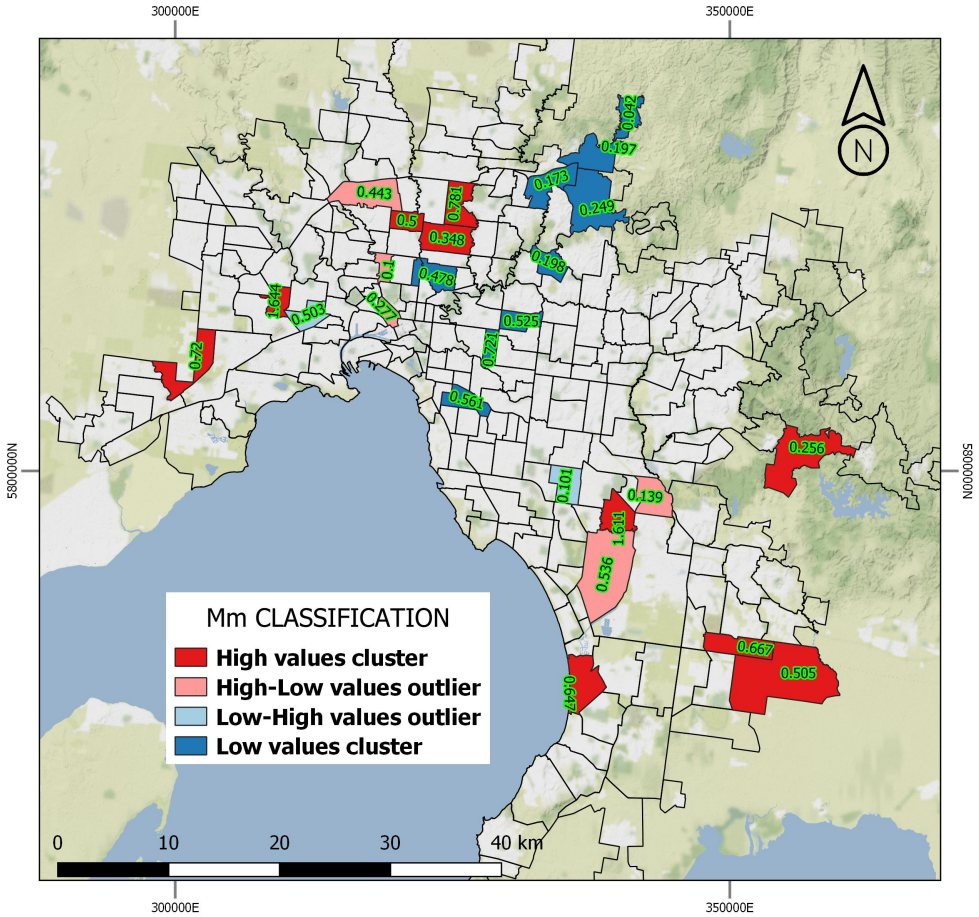


Figure 6.7: Multivariate clusters and outliers classification according to the Mm_c Mm_o measures. Labels include the absolute values of the computed Mm_c or Mm_o at each significant location. *Basemap: ©OpenStreetMap contributors, ©Stamen Design.*

6.2.1 Results disclosure and validation

The significant clustering outlined in Figure 6.4, suggests the existence of *interesting* locations, and in turn non-random patterns, within the reciprocal spatial arrangement of the considered indices. The multivariate Local Geary's c map provides a snapshot of spatial association patterns which are not apparent by observing the index maps independently (Figure 6.2). An assessment of the reliability of both clusters and outliers detection is achieved by introducing the D_i measure (Figure 6.5). The clusters and outliers classification is performed by means of the Mm indicator (Figure 6.7).

Local clustering is numerically described using the computed D_i values in terms of attributes dispersion in the PC plane. Focusing on cluster locations, the smaller dispersions are concentrated South-East to the Central Business District (CBD). These locations are classified as low values clusters according to the Mm indicator. Hence, higher performances in terms of social vulnerability are expected at these locations. High values clusters are more scattered and generally characterised by a higher dispersion. They are located mainly North to the CBD as well as along the South-Eastern and Western borders of the Greater Melbourne region. Focusing on outliers, these are detected South-West to the CBD as well as along the South-Eastern edge of the Greater Melbourne region. Their position generally corresponds to the transition area between clusters and not significant locations hence their presence may be connected to this finding. No further conclusions are here argued about both the outliers meaning in terms of social vulnerability as well as the practical results application into any specific urban management practice. Nevertheless, the experiment demonstrates the applicability of the procedure on tradition spatial data providing asset maps that might support needs assessments, policies evaluation, and interventions planning that target social vulnerability reduction.

Profile indicator	**Description, [Unit of measure]**
Labour ratio	Number of labourers divided by the number of people in the labor force, [%].
Tertiary education ratio	Number of people having completed the highest level of education divided by the total resident people, [%].
Low-income ratio	Number of people having a total personal income lower than 599 AUD$/week divided by the total resident people, [%].

Table 6.4: Vulnerability profile indicators considered in the validation of multivariate spatial association mapping results.

Obtained results can be validated through comparisons with reference data that consist of different information targeting social vulnerability than the one adopted in the spatial association analysis. These may include vulnerability profiles which are traditionally developed and used for a-spatial analyses. Profiles are generally composed by a set of indicators for each district or parcel derived from census and socio-economical data. In this case study, three profile indicators are considered for validation purposes. These are the labour ratio, the tertiary education ratio, and the low-income ratio computed for the reference Year 2011 as described in Table 6.4[3]. Comparison between the three indicators is carried out on the detected clusters and outliers in Figure 6.8.

The comparison is disclosed according to the meaning of both indicators and high/low values clusters in terms of social vulnerability. Indeed, it is expected to observe relative lower trends for low-income and labour ratios as well as a relatively higher trend of the tertiary education ratio at low values (i.e. LL) cluster locations. Vice-versa at high values cluster locations (i.e. HH). As shown in Figure 6.8, the expectation is met especially looking at the trend of labour and the tertiary education ratios. The low-income ratio shows a less marked agreement with the expectations. Nevertheless, the highest peaks are registered

[3]Indicators for validation were kindly provided by the Centre for Spatial Data Infrastructures and Land Administration (CSDILA) of the University of Melbourne (http://www.csdila.unimelb.edu.au)

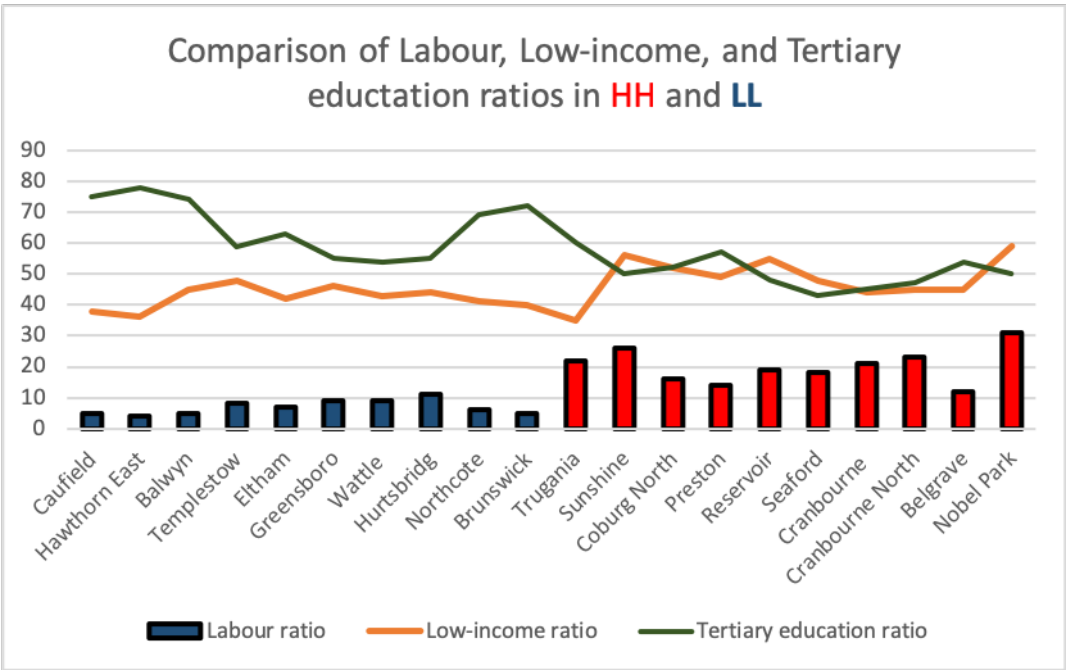


Figure 6.8: Trends of the considered vulnerability profile indicators at each low values (blue bar) and high values (red bar) cluster location.

within the high values clusters. Despite being the latter a preliminary and coarse validation procedure, results provide additional evidence on the reliability of the multivariate spatial association mapping outcomes. Therefore, multivariate LISA maps are promising tools to support operational urban management such as locating and understanding critical social vulnerability patterns.

6.2.2 Discussion

The aim of this closing section is to outline and remark the main features of the proposed procedure by also including the lesson learned through the development of the preliminary case study.

The multivariate Local Geary's c demonstrates to be suitable for mapping multivariate spatial association patterns. A reasonable classification technique for cluster and outlier types is achieved in a reproducible way by applying medians and means comparison, i.e through the Mm measure. Robustness and reli-

ability of the Mm classification require additional investigation. Considering multiple variables, the centrality measures might be affected by local anomalies in the variable values such as points spanning around the mean (or median) thus smoothing or leaping the differences on which the Mm indicator is based. However, this classification can be easily fit into a single analysis pipeline together with the multivariate Local Geary's c, allowing for an agile integration into a single software tool, e.g. a QGIS plugin. As a proof of the above, the prototype Python code used in the presented case study has been made available on the GitHub[4].

An additional measure of multivariate spatial association is outlined, i.e. the D_i and computed considering the original variable projection onto a reduced attribute space through PCA. Results from the D_i are aligned with the outcomes of the Local Geary's c. The use of PCs is here introduced for better targeting analyses involving a large number of variables thus like a simple dimensionality reduction strategy. The contribution of PCA has not been exhaustively discussed in the case study. This is partially due to the limited dimension of the considered data which can be still partially explored by means visual analysis of choropleth maps. By increasing the number of variables, the PCA might become more useful in order to reduce the data dimension as well as to enable a more compact results visualisation. In principle, the D_i might be directly computed also on the original standardized variables. Both significance testing and clusters/outliers detection might be carried out using the same strategies that are used for the multivariate Local Geary's c, namely by means of conditional permutations under CSR hypothesis and comparisons with the expectation. An early method proposal for adapting the D_i into a local indicator of multivariate spatial association is included in Appendix B.

[4]https://github.com/danioxoli/multivar_lisa

Optimal settings for the multivariate LISA mapping, e.g. the significance level, need for further investigation to properly account for a larger number of variables as well as for the possible effects of the spatial association affecting every single variable. Therefore at this early stage, results have to be intended as experimental, and the robustness of the procedure requires for additional tests.

Chapter 7

Conclusions

The concept of geospatial data characters with a focus on the spatial association - together with novel methods and tools to explore them - have been presented. The crucial role still covered by the spatial association in the modern geospatial data analysis has been most discussed and investigated by leveraging ESDA. In particular, the use of ESDA within FOSS GIS environments has been argued along this work. This has been identified among the best practices to produce snapshots of spatial phenomena at different scales, to highlight their underlying spatial features, and to guide the analysts in data assessments and interpretations. Moreover, the inclusion of ESDA modules into the most popular FOSS GIS - such as QGIS - has been identified as a meaningful objective to enforce ESDA use among a broader and diverse users' community. I firmly believe that to pursue such a goal, the most effective way is advocating openness into all the edges of science. This by starting from source code, passing through data and models that are at the base of any advanced knowledge and rational awareness.

Despite software technologies today offer massive support to any application, there is always a need for extending, revamping, or re-designing established geospatial analysis methods and frameworks. This to unpin the real assets of

the emerging geospatial data to meet urgent requirements and expectations associated with them by users. With this in mind, early extensions and complements of LISA methods to the multivariate context have been proposed. The underlying goal is to address frontier challenges of spatial statistics and geo-visualisation that include - among others - multivariate spatial patterns detection and mapping [48]. The latter is promising for many disciplines that require simultaneous explorations of multiple variables to be linked with any complex natural or human phenomenon under investigation. These include subjects such as disaster risk management, ecology, epidemiology, regional and social science among others.

Extensive testing of the proposed methodologies has not being carried out within this work. This leave places for further research that should mainly focus on investigating the analytical implications of multivariate spatial association into the confirmatory analysis.

Future directions for the work will then focus on additional validations of the proposed methodology for multivariate spatial clusters and outliers classification (Mm_i) as well as on the assessment of the statistical validity of the D_i measure (Section 6.1). An early test on the latter has been set up in Appendix B. Concerning the software side, the source code produced within this work requires a substantial improvement to be integrated into a plugin for the most popular GIS platforms, such as QGIS. The extension of LISA to the multivariate context implies higher computational costs due to the concurrent analysis of multiple variables. The introduction of parallel computing to cut down the multivariate LISA computational time is advised due to the critical role of this factor to the practical application of this technique. The same applies by considering the analysis of high-resolution datasets for large geographic regions that is one of the frontiers of the modern geospatial data analysis.

Although this work represents only a byte of the modern geospatial analysis solutions, I believe that all those disciplines which have made proficient use of LISA in the past, would take advantage in the future for the multivariate LISA analysis to address new challenges of the *Big Data* era.

Appendix A

The Hotspot Analysis Plugin Functionalities

Functionality	**Description**
LISA	• Getis-Ord G_i^* • Local Moran's I (univariate and bivariate)
INPUT DATA	A shapefile of points or polygons with a projected CRS and at least one numerical attribute at each geometry.
ATTRIBUTE SELECTION	Field name of the shapefile attribute table containing the analysis variable. Double selection for the bivariate Local Moran's I.

SPATIAL WEIGHTS MATRIX	• Fixed-distance band. Default distance is the one ensuring at least one neighbour to each location. The user can specify a distance in the the unit of measure of the input data CRS (default for points) • First order Queen's case contiguity matrix (default for polygons) • Optimized fixed-distance band such as the one maximizing the Global Moran's I for the dataset, selected among a user defined range (alternative for points only) • KNN spatial weights with a user selected k (alternative for points only) • Row standardized weights (alternative to binary weights both for points and polygons)
SIGNIFICANCE TEST	• Normality approximation (default) • Conditional permutations (alternative)
OUTPUT DATA	A copy of the input shapefile with two new fields in the attribute table, i.e. LISA Z-scores and pseudo p-values.
MAPPPING	Automatic styling, rendering and legend creation for the output layer on QGIS. Default hues are given by combining Z-scores and pseudo p-values.

Table A.1: The Hotspot Analysis Plugin functionalities.

Appendix B

The D_i as a Local Indicator of Multivariate Spatial Association: An Early Method Proposal

Chapter 6 introduces the local multivariate measure of dispersion D_i which consists of an average distance in a k dimensional attribute space from observations at a focal location i and at its spatial neighbours j to their centre of mass (or centroid). The centre of mass is the mean position in the attribute space computed from all the k observations at each considered location and its neighbours. The D_i is initially applied to a reduced attribute space (bi-dimensional) obtained by means of PCA. Under these circumstances, benefits of the D_i measure into the practical exploration of multivariate spatial association are illustrated through an application on a real case study (Section 6.2). However, the D_i can be easily adapted to k dimensional spatial data that to a general multivariate context where realisations of k spatial variables are observed at each location of the study region. This appendix aims to argue the theoretical applicability the D_i method whereas testing and critical reviews are places left for future researches.

The D_i measure can be generalised to produce a local indicator of multivariate spatial association as follows. Let's consider a generic spatial dataset composed of n locations to which z_k observations of the same k variables are associated. Observations are expressed in deviation from the mean through Z-score scaling (Equation 3.6). Spatial relationships are defined by means of a binary spatial weights matrix W^b having non zero diagonal such as $W^b_{i,i} = 1$. The purpose is to compute the D_i at each location and to perform inference under CSR hypothesis. This can be achieved by using the same strategy such as for other LISA, i.e. using conditional permutations. The test statistic is outlined in Equation B.1.

$$D_i = \frac{\sum_{j=1}^{n} W^b_{i,j} \cdot \sqrt{\sum_{v=1}^{k} (\bar{z}_{v,i} - z_{v,j})^2}}{\sum_{j=1}^{n} W^b_{i,j}}; \bar{z}_{v,i} = \frac{\sum_{j=1}^{n} W^b_{i,j} \cdot z_{j,v}}{\sum_{j=1}^{n} W^b_{i,j}}; i \in j \tag{B.1}$$

$W^b_{i,j}$ is the element of W^b for each couple i and j. $\bar{z}_{v,i}$ is the centre of mass coordinate on the v axis of the k dimensional attribute space with $v \in k$. The centre of mass is computed out of the locations subset having a non zero entry in the row vector W^b_i that is the vector of weights of the focal location i. $z_{v,j}$ is the standardized observation of the variable v at location j. A value of the D_i can be now computed at each location in the dataset. Significance under CSR is inferred by means of conditional permutations, therefore by holding the vector of observations at location i fixed, random permute or shuffle the remaining observations across the other $n - 1$ locations, and recompute the local test statistic. By repeating m times this process, an empiric reference distribution for the statistic under the CSR hypothesis is obtained. Pseudo p-values can be then computed as for other LISA (Equation 3.3). The computed D_i can be compared either with its sample mean or the mean from permutations to clas-

sify significant locations into clusters or outlier like for the multivariate Local Geary's c ($c_{k,i}$).

The D_i and the $c_{k,i}$ embed a similar concept of the spatial association that is measured by means of dissimilarities rather than correlation among observations. Dissimilarities are expressed in terms of a weighted sum of squared distances in the attribute space by the $c_{k,i}$, whereas through average distances or weighted root-mean-square differences from a centre of mass by the D_i. The centre of mass is here considered as a reference for each locations subset composed of i location and its neighbours. This is the main difference from the $c_{k,i}$ that adopts i as the reference to compute square distances. The use of a centre of mass instead of the focal location i provides with a slightly different concept of spatial association. Generally speaking, the $c_{k,i}$ measures how far are observations of the k variables between location i and its neighbours j. The D_i instead produces a measure of dispersion around a virtual point (centre of mass). In principle, this is the most representative position in the k dimensional attribute space for the locations subset which is a candidate spatial cluster or outliers. The D_i attempts to summarise the dissimilarity of observations in the locations subset from this representative position. With this in mind, the smaller the dissimilarity, the higher the chance to encounter a cluster, therefore positive spatial association, vice-versa for an outlier. The definition of both *smaller* and *higher* always refers to the relative range of the D_i values for the considers dataset. The focal location i loses its central role in the definition of the spatial association while the focus is on the choral contribution of the locations subset by outlining the chance of being properly represented by a single position in the attribute space.

A preliminary comparison between the outcomes of D_i and the $c_{k,i}$ is carried out using the Guerry's data on moral statistics of France in the 1830s [48]. This

data includes six variables for 85 France departments. Corsica (an island) is removed from the original dataset. Both the definition and the spatial distribution of the variables are included in Figure B.1a. The same dataset was also used in the original publication of the multivariate Local Geary's c [8]. Indeed, the parameters set-up proposed by [8] for the $c_{k,i}$ test is here adopted for the D_i to enable the comparison. Namely are considered a queen's contiguity spatial weights matrix[1], a significance level $\alpha = 0.01$, a number of permutations $m = 99999$, and the FDR correction by means of the Benjamini-Hochberg procedure [10]. The comparison between the $c_{k,i}$ and the D_i maps is included in Figure B.1b. Twenty-one significant locations are spotted using the $c_{k,i}$ while 26 using the D_i, of which 17 overlaps. Hence, multivariate patterns of spatial association are comparable for the two maps nevertheless not identical due to the slightly different approaches that are used by the two statistics for describing spatial association. The outperforming of the D_i should be investigated starting from the use of a centre of mass rather than a focal location for computing dissimilarities. This may dilute the spatial association characteristics among neighbouring locations thus smoothing or averaging patterns across relatively larger sub-regions.

In views of the above, D_i is here proposed as a complementary local indicator of multivariate spatial association to the $c_{k,i}$. Indeed, the D_i provides a slightly different measure of spatial association that can be employed together with the $c_{k,i}$ to provide additional power against the CSR hypothesis. Furthermore, the D_i values computed from standardized observations give a metric of the local clustering which might be used to correct multivariate spatial models for local effects as well as into generic exploratory experiments.

For a matter of simplicity, a binary spatial weights matrix W^b has been consid-

[1]Notice that differences between the two maps in Figure B.1b may be partially due to the use of a binary spatial weights matrix for computing the D_i according to Equation B.1 instead of a row-standardized matrix. This particular could not be fully ascertained from [8].

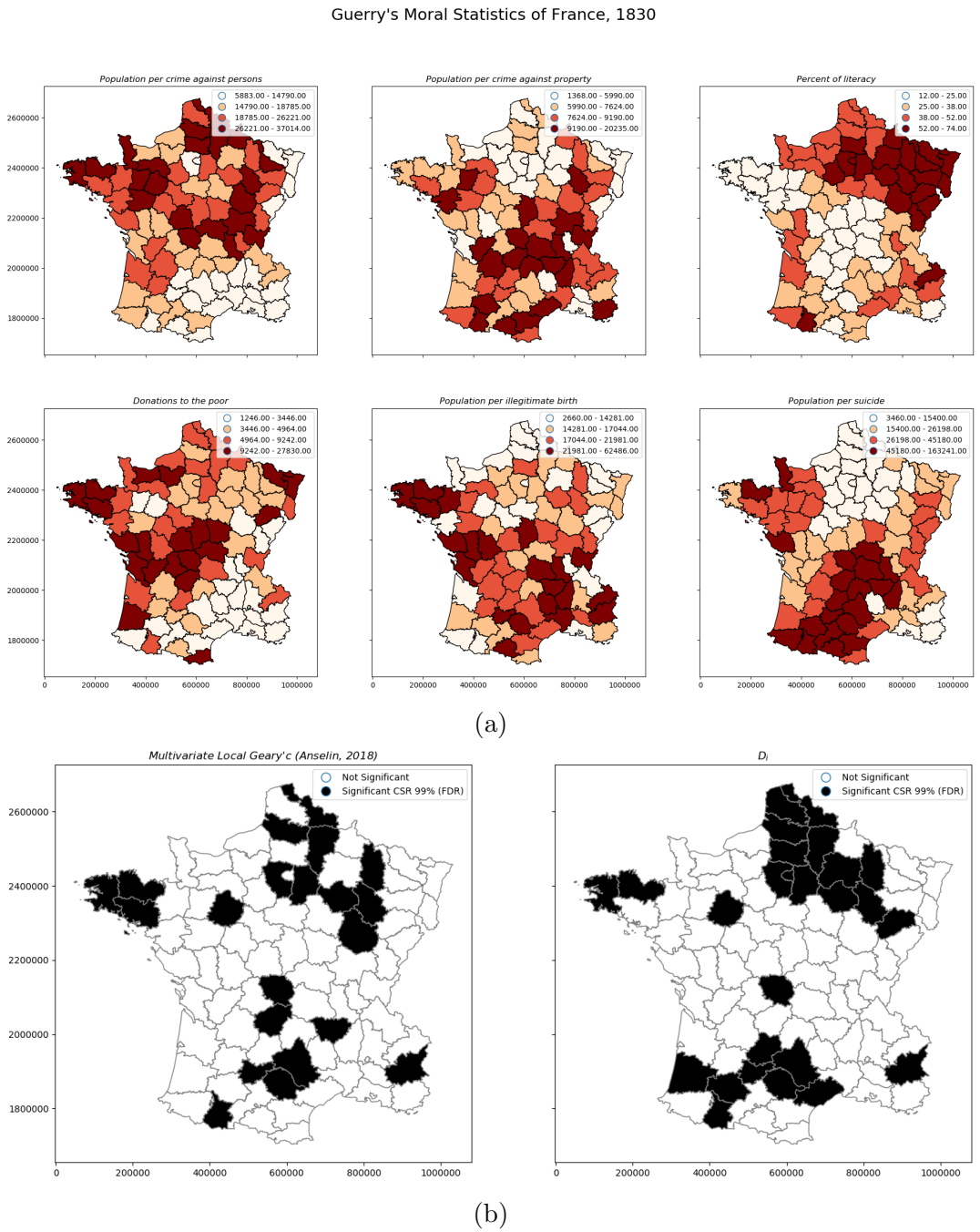


Figure B.1: (**a**) Quantile maps of the six Guerrys moral statistics of France. (**b**) multivariate Local Geary's c map (right) published in [8] and the computed D_i map (left).

ered for the D_i computations so far. This implies that both the centre of mass coordinates as well as the average distances are considered as simple arithmetic mean of observations in the locations subset considered, that is composed of the i location and its neighbours j. However, spatially weighted averages can be introduced either for computing the centre of mass coordinates and the av-

erage distances (or both) by adopting different types of spatial weights. These include, e.g. distance decay spatial weights as well as row-standardized weights among others. Moreover, the influence of every variable on the multivariate patterns can be arbitrarily modulated by introducing variables weights either in computing the centre of mass coordinates and the average distances (or both). Custom spatial and variables weighting may unpin valuable tuning options at the cost of introducing additional complexity to the interpretations of results. Equation B.2 present the generalized D_i test statistic to account for custom weightings, where W is a generic spatial weights matrix while a_v and b_v are the variable weights. This topic is here intended as a simple proposal that requires further investigations which have not been addressed in the present study.

$$D_i = \frac{\sum_{j=1}^{n} W_{i,j} \cdot \sqrt{\sum_{v=1}^{k} \cdot b_v \cdot (\bar{z}_{v,i} - z_{v,j})^2}}{\sum_{j=1}^{n} W_{i,j}}; \bar{z}_{v,i} = \frac{\sum_{j=1}^{n} W_{i,j} \cdot a_v \cdot z_{j,v}}{\sum_{j=1}^{n} W_{i,j}}; i \in j \qquad \text{(B.2)}$$

Due to the experimental content of the multivariate methods discussed above, the following general recommendations have to be stated. As also argued by [8], the introduction of multiple dimensions in the spatial association analysis potentially produces trade-offs among spatial variables leading to patterns which cannot be disclosed by merely overlaying the univariate patterns of each variable. Therefore, dimensionality may negatively affect the usefulness of the analysis by driving to blind alleys in results interpretation. A critical review and validation of both results and analysis strategy adopted are strongly suggested within any practical application.

List of Acronyms

IT Information Technology

EO Earth Observations

ESDA Exploratory Spatial Data Analysis

FOSS Free and Open Source Software

GIS Geographic Information Systems

LISA Local Indicators of Spatial Association

VGI Volunteered Geographic Information

EDA Exploratory Data Analysis

CRS Coordinate Reference System

GWR Geographically Weighted Regression

CSR Complete Spatial Randomness

FDR False Discovery Rate

GPL GNU General Public License

OS Operative Systems

KML Keyhole Markup Language

GDAL Geospatial Data Abstraction Library

API Application Programming Interfaces

GUI Graphical User Interface

spdep Spatial Dependence: Weighting Schemes, Statistics and Models

PySAL Python Spatial Analysis Library

BSD Berkeley Software Distribution

STARS Space-Time Analysis of Regional Systems

CSV Comma-separated Values

GAL GenePix Array List

DBF DataBase File

OSGeo Open Source Geospatial Foundation

SNAP Sentinel Application Platform

PyQGIS Python QGIS

VCS Version Control Systems

KNN K-Nearest Neighbours

GNSS Global Navigation Satellite Systems

GPX GPS eXchange Format

ISPRA Italian National Institute for Environmental Protection and Research

CC-BY Creative Commons

PCA Principal Component Analysis

PC Principal Components

VAMPIRE Vulnerability Assessment for Mortgage, Petrol and Inflation Risks and Expenditure

IRSD Index of Relative Socio-economic Disadvantage

SA2 Statistical Areas Level 2

ASGS Australian Statistical Geography Standard

ABS Australian Bureau of Statistics

OECD Organization for Economic Co-operation and Development

CBD Central Business District

CSDILA Centre for Spatial Data Infrastructures and Land Administration

Bibliography

[1] Adams, B., McKenzie, G., Gahegan, M. (2015). Frankenplace: interactive thematic mapping for ad hoc exploratory search. In: *Proceedings of the 24th international conference on world wide web*, 12-22. International World Wide Web Conferences Steering Committee.

[2] Anselin, L., Griffith, D. A. (1988). Do spatial effects really matter in regression analysis?. *Papers in Regional Science*, 65(1), 11-34.

[3] Anselin, L., Rey, S. (1991). Properties of tests for spatial dependence in linear regression models. *Geographical analysis*, 23(2), 112-131.

[4] Anselin, L. (1995). Local indicators of spatial association LISA. *Geographical analysis*, 27(2), 93-115.

[5] Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. *Geographical Information Systems: principles, techniques, management and applications*, 1, 251-264.

[6] Anselin, L., Syabri, I., Kho, Y. (2006). GeoDa: an introduction to spatial data analysis. *Geographical analysis*, 38(1), 5-22.

[7] Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in regional science*, 89(1), 3-25.

[8] Anselin, L. (2018). A Local Indicator of Multivariate Spatial Association: Extending Geary's c. *Geographical analysis*.

[9] Bartels, C. P., Ketellapper, R. H. (1979). *Exploratory and explanatory statistical analysis of spatial data*. Boston: Martinus Nijhoff.

[10] Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society*. Series B (Methodological), 289-300.

[11] Brovelli, M. A., Oxoli, D., Zurbaràn, M. A. (2016). Sensing slow mobility and interesting locations for Lombardy Region (Italy): A case study using pointwise geolocated open data. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41, 603-607.

[12] Besag, J., Diggle, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Applied statistics*, 327-333.

[13] Bivand, R. S. (2010). Exploratory spatial data analysis. In: *Handbook of applied spatial analysis*, 219-254. Springer, Berlin, Heidelberg.

[14] Bolon, B. R., Ghabra, J. A., Ward, M. L. (2014) *U.S. Patent No. 8,849,254.* Washington, DC: U.S. Patent and Trademark Office.

[15] Boots, B. (2002). Local measures of spatial association. *Ecoscience*, 9(2), 168-176.

[16] Brunsdon, C., Fotheringham, A. S., Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4), 281-298.

[17] Buckley, A. (2013). *Understanding statistical data for mapping purposes.* ArcUser Winter 2013. ESRI.

[18] Campelo, M. B., Mendes, R. M. N. (2014). Comparing Webshare services to assess MTB use in protected areas. In: *Proceedings of the 7th International Conference on Monitoring and Management of Visitors in Recreational and Protected Areas*, 161-163.

[19] Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., Purves, R. (2016). *European handbook of crowdsourced geographic information.* Ubiquity Press.

[20] Chen, W., Guo, F., Wang, F. Y. (2015). A survey of traffic data visualization. *IEEE Transactions on Intelligent Transportation Systems*, 16(6), 2970-2984.

[21] Chica-Olmo, J. (2007). Prediction of housing location price by a multivariate spatial method: cokriging. *Journal of Real Estate Research*, 29(1), 91-114.

[22] Chipman, J., Wright, R., Ellis, M., Holloway, S. R. (2012). Mapping the evolution of racially mixed and segregated neighborhoods in Chicago. *Journal of Maps*, 8(4), 340-343.

[23] Cliff, A., Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical analysis*, 4(3), 267-284.

[24] Cliff, A. D., Ord, J. K. (1973) Spatial autocorrelation. *Pion*, London.

[25] Congedo, L., La Mantia, C., Luti, T., Marinosci, I., Raudner, A., Riitano, N., Strollo, A., Garofalo, V., Mastrorosa, S., Meccoli, L., Rossi, L., Vitaletti, A., Munaf M. (2016) Stima del consumo di suolo a livello provinciale e comunale. In: *Consumo di suolo, dinamiche territoriali e servizi ecosistemici. Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA)*, Rapporti 248/2016.

[26] Cressie, N. (1990). The origins of kriging. *Mathematical geology*, 22(3), 239-252.

[27] Crompvoets, J., De Man, E., Geudens, T. (2010). Value of spatial data: networked performance beyond economic rhetoric. *IJSDIR*, 5, 96-119.

[28] Cutter, S. L. (2003). GI science, disasters, and emergency management. *Transactions in GIS*, 7(4), 439-446.

[29] Daly, C., Neilson, R. P., Phillips, D. L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of applied meteorology*, 33(2), 140-158.

[30] Dark, S. J., Bram, D. (2007). The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography*, 31(5), 471-479.

[31] De Knegt, H. J., van Langevelde, F. V., Coughenour, M. B., Skidmore, A. K., De Boer, W. F., Heitknig, I. M. A., Knox, N. M., Slotow, R., van der Waal, C., Prins, H. H. T. (2010). Spatial autocorrelation and the scaling of species-environment relationships. *Ecology*, 91(8), 2455-2465.

[32] Di Leginio, M., Fumanti, F., Strollo, A., Munaf, M. (2016). Funzioni del suolo, servizi ecosistemici e minacce. In: Consumo di suolo, dinamiche territoriali e servizi ecosistemici. *Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA)*, Rapporti 248/2016.

[33] Demar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., McLoone, S. (2013). Principal component analysis on spatial data: an overview. *Annals of the Association of American Geographers*, 103(1), 106-128.

[34] Dent, B. D. (1972). A note on the importance of shape in cartogram communication. *Journal of Geography*, 71(7), 393-401

[35] Dessers, E., Crompvoets, J., Janssen, K., Vancauwenberghe, G., Vandenbroucke, D., Vanhaverbeke, L., Van Hootegem, G. (2012). A multidisciplinary research framework for analysing the spatial enablement of public sector processes. *IJSDIR*, 7, 125-150.

[36] Dodge, Y. (2006). *The Oxford dictionary of statistical terms*. New York: Oxford University Press Inc.

[37] Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global ecology and biogeography*, 16(2), 129-138.

[38] Dubes, R., Jain, A. K. (1980). Clustering methodologies in exploratory data analysis. In:*Advances in computers*, 19, 113-228. Elsevier.

[39] Dubin, R. A. (1998). Spatial autocorrelation: a primer. *Journal of housing economics*, 7(4), 304-327.

[40] Efron, B., Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[41] Efron, B., Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press.

[42] Fearn, T., Thompson, M. (2001). A new test for "sufficient homogeneity". *Analyst*, 126(8), 1414-1417.

[43] Fischer, M. M., Getis, A. (2009). *Handbook of applied spatial analysis: software tools, methods and applications*. Springer Science & Business Media.

[44] Fotheringham, A. S. (1981). Spatial structure and distance-decay parameters. *Annals of the Association of American Geographers*, 71(3), 425-436.

[45] Fotheringham, A. S. (1997). Trends in quantitative methods I: stressing the local. *Progress in Human Geography*, 21(1), 88-96.

[46] Fotheringham, A. S. (2009). "The problem of spatial autocorrelation and local spatial statistics. *Geographical analysis*, 41(4), 398-403.

[47] Franke, B., Plante, J.F., Roscher, R., Lee, E.S.A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M.M. (2016). Statistical inference, learning and models in big data. *International Statistical Review*, 84(3), 371-389.

[48] Friendly, M. (2007). A.-M. Guerry's" Moral Statistics of France": Challenges for Multivariable Spatial Analysis. *Statistical Science*, 368-399.

[49] Gardner, N. (2009). A manifesto for slow travel. *Hidden Europe Magazine*, 25(1), 14.

[50] Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3), 115-146.

[51] Getis, A., Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3), 189-206.

[52] Getis, A., Ord, J. K. (1996). Local spatial statistics: an overview. *Spatial analysis: modelling in a GIS environment*, 374, 261-277.

[53] Getis, A., Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical analysis*, 36(2), 90-104.

[54] Getis, A. (2007). Reflections on spatial autocorrelation. *Regional Science and Urban Economics*, 37(4), 491-496.

[55] Getis, A. (2010). Spatial Autocorrelation. In: *Handbook of applied spatial analysis*, 255-278. Springer, Berlin, Heidelberg.

[56] Gilani, H. (2005). Automatically Determining Route and Mode of Transport Using a GPS Enabled Phone. PhD thesis, *University of South Florida*.

[57] Goodman, R. (2017). Melbourne: Growing pains for the liveable city. In: *Planning Metropolitan Australia*, 59-83, Routledge.

[58] Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

[59] Goodchild, M. F. (1986). Spatial autocorrelation. *Geo Books*, 47.

[60] Graser, A., Olaya, V. (2015). Processing: A python framework for the seamless integration of geoprocessing tools in QGIS. *ISPRS International Journal of Geo-Information*, 4(4), 2219-2245.

[61] Griffith, D. A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. *Annals of the Association of American Geographers*, 95(4), 740-760.

[62] Hahmann, S., Burghardt, D. (2013). How much information is geospatially referenced? Networks and cognition. *International Journal of Geographical Information Science*, 27(6), 1171-1189.

[63] Haining, R. (1977). Model specification in stationary random fields. *Geographical Analysis*, 9(2), 107-129.

[64] Hay, S. I., George, D. B., Moyes, C. L., Brownstein, J. S. (2013). Big data opportunities for global infectious disease surveillance. *PLoS medicine*, 10(4), e1001413.

[65] Hey, T., Tansley, S., Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery*, 1. Redmond, WA: Microsoft research.

[66] Hardy, D., Frew, J., Goodchild, M. F. (2012). Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science*, 26(7), 1191-1212.

[67] Hesse, B. W., Moser, R. P., Riley, W. T. (2015). From big data to knowledge in the social sciences. *The Annals of the American Academy of Political and Social Science*, 659(1), 16-32.

[68] Jiang, B. (2013). Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer*, 65(3), 482-494.

[69] Jiang, B. (2018). Spatial Heterogeneity, Scale, Data Character and Sustainable Transport in the Big Data Era. *ISPRS Int. J. Geo-Inf.*, 7(5), 167.

[70] Jolliffe, I. (2011). Principal component analysis. In: *International encyclopedia of statistical science*, 1094-1096. Springer, Berlin, Heidelberg.

[71] Kelejian, H. H., Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1), 53-67.

[72] Krakk, M., Ormeling, F. (2003). *Cartography: Visualization of Geospatial Data*. Pearson Education Limited, Edinburgh Gate.

[73] La Rocca, R. A. (2010). Soft mobility and urban transformation. Tema. *Journal of Land Use, Mobility and Environment*, 2.

[74] Lawhead, J. (2015). *QGIS python programming cookbook*. Packt Publishing Ltd.

[75] Lee, S. I. (2001). Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I. *Journal of geographical systems*, 3(4), 369-385.

[76] Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm?. *Ecology*, 74(6), 1659-1673.

[77] Leung, Y., Mei, C. L., Zhang, W. X. (2003). Statistical test for local patterns of spatial association. *Environment and Planning A*, 35(4), 725-744.

[78] Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 119-133.

[79] Liu, J., Dietz, T., Carpenter, S. R., Folke, C., Alberti, M., Redman, C. L., Schneider, S. H., Ostrom, E., Pell, A. N., Lubchenco, J., Taylor, W. W., Ouyang, Z., Deadman, P., Kratz, T., Provencher, W. (2007). Coupled human and natural systems. *AMBIO: a journal of the human environment*, 36(8), 639-649.

[80] Lloyd, C. D., Lilley, K. D. (2009). Cartographic veracity in medieval mapping: analyzing geographical variation in the Gough Map of Great Britain. *Annals of the Association of American Geographers* 99(1), 27-48.

[81] Lloyd, C. D. (2010). *Local models for spatial analysis*. CRC press.

[82] Menke, K., Smith Jr, R., Pirelli, L., Van Hoesen, J., Davis, P. (2015). *Mastering QGIS*. Birmingham, UK: Packt Publishing.

[83] Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17-23.

[84] Morelli F. (2014). *Il reddito degli italiani in ogni Comune, le disuguaglianze e dove vivono i super-ricchi*. Available online: http://www.opendatabassaromagna.it/2014/04/il-reddito-degli-italiani-in-ogni.html (accessed 15/12/17).

[85] Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T., Meehl, G. A., Mitchell, J. F. B., Nakicenovic, N., Riahi, K., Smith, S. J., Stouffer, R. J., Thomson, A. M., Weyant, J. P., Wilbanks, T. J. (2010). The next generation of scenarios for climate change research and assessment. *Nature*, 463(7282), 747-756.

[86] Lwe, P., Neteler, M. G. (2014). Data science: history repeated?: the heritage of the free and open source GIS community. In: *EGU General Assembly*.

[87] Ord, J. K., Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, 27(4), 286-306.

[88] Ord, J. K. (2004). Spatial processes. In: *Encyclopedia of Statistical Sciences*, 12, 7889-7890.

[89] Oxoli, D., Prestifilippo, G., Bertocchi, D., Zurbarn, M. (2017). Enabling spatial autocorrelation mapping in QGIS: The Hotspot Analysis Plugin. *GEAM. GEOINGEGNERIA AMBIENTALE E MINERARIA*, 151(2), 45-50.

[90] Oxoli, D., Molinari, M.E., Brovelli, M.A (2018). Hotspot Analysis, an open source GIS tool for exploratory spatial data analysis: application to the study of soil consumption in Italy. *Rend. Online Soc. Geol. It.*, 46, 82-87.

[91] Osaragi, T. (2002) *Classification methods for spatial data representation.* CASA Working Papers 40. Centre for Advanced Spatial Analysis (UCL): London, UK.

[92] Phipson, B., Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1).

[93] Piry, S., Chapuis, M. P., Gauffre, B., Papax, J., Cruaud, A., Berthier, K. (2016). Mapping Averaged Pairwise Information (MAPI): a new exploratory tool to uncover spatial structure. *Methods in Ecology and Evolution*, 7(12), 1463-1475.

[94] Rajsingh, E. B., Veerasamy, J., Alavi, A. H., Peter, J. D. (Eds.). *Advances in Big Data and Cloud Computing.* Springer Nature Singapore Pte Ltd., 2018.

[95] Ramadani, V., Zendeli, D., Gerguri-Rashiti, S., Dana, L. P. (2018). Impact of geomarketing and location determinants on business development and decision making. Competitiveness Review: *An International Business Journal*, 28(1), 98-120.

[96] Riitano, N., Congedo, L., Garofalo, V., La Mantia, C., Luti, T., Marinosci, I., Mastrorosa, S., Meccoli, L., Raudner, A., Rossi, L., Strollo, A., Vitaletti, A., Munaf, M. (2016). Stima del consumo di suolo a livello nazionale e regionale. In: *Consumo di suolo, dinamiche territoriali e servizi ecosistemici.* Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA), Rapporti 248/2016.

[97] Smith, D. A. (2016). Online interactive thematic mapping: Applications and techniques for socio-economic research. *Computers, Environment and Urban Systems*, 57, 106-117.

[98] Sokal, R. R., Oden, N. L., Thomson, B. A. (1998). Local spatial autocorrelation in a biological model. *Geographical Analysis*, 30(4), 331-354.

[99] Steudler, D., Rajabifard, A. (Eds.). *Spatially enabled society.* International Federation of Surveyors, 2012.

[100] Tate, E. (2012). Social vulnerability indices: a comparative assessment using uncertainty and sensitivity analysis. *Natural Hazards*, 63(2), 325-347.

[101] Taubenbck, H., Esch, T., Felbier, A., Wiesner, M., Roth, A., Dech, S. (2012). Monitoring urbanization in mega cities from space. *Remote sensing of Environment*, 117, 162-176.

[102] Tiefelsdorf, M. (2002). The saddlepoint approximation of Moran's I's and local Moran's Ii's reference distributions and their numerical evaluation. *Geographical Analysis*, 34(3), 187-206.

[103] Tobler, W., (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234-240.

[104] Tobler, W. R. (1973). Choropleth maps without class intervals?. *Geographical analysis*, 5(3), 262-265.

[105] Tukey, J. W. (1977). *Exploratory Data Analysis*. Pearson.

[106] Van der Zee, E., Bertocchi, D., Janusz, K. (2016). Using Big Data to discover how the maturity of a heritage destination influences the use and attractiveness of urban cultural landscape. A case study of Antwerp, Bolzano and Krakw. In: *Proceedings of TCL2016 conference: Tourism and cultural landscapes: Towards a sustainable approach*, 614-628.

[107] Van Der Aalst, W. M., La Rosa, M., Santoro, F. M. (2016). Business process management. *Bus. Inform. Syst. Eng.*, 58(1), 1-6.

[108] Velleman, P. F., Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Duxbury Press.

[109] Von Hippel, P. T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education*, 13(2).

[110] Walter, S. D. (1992). The analysis of regional patterns in health data: II. The power to detect environmental effects. *American Journal of Epidemiology*, 136(6), 742-759.

[111] Wartenberg, D. (1985). Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, 17(4), 263-283.

[112] Wong, D. W. (1997). Spatial dependency of segregation indices. *Canadian Geographer*, 41(2), 128-136.

[113] Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between Time-Series?–a study in sampling and the nature of time-series. *Journal of the royal statistical society*, 89(1), 1-63.

[114] Zhang, J., You, S., Gruenwald, L. (2017). Towards GPU-Accelerated Web-GIS for Query-Driven Visual Exploration. In: *International Symposium on Web and Wireless Geographical Information Systems*, 119-136.

[115] Zurbaràn, M., Wightman, P. M., Brovelli, M. A,. Oxoli, D., Iliffe, M., Jimeno, M., Salazar, A. (2018). NRand-K: Minimizing the Impact of Noise-Based Location Obfuscation in Spatial Analysis. *Transactions in GIS*, 22(5), 1257-1274.