RESEARCH ARTICLE

A classification technique for local multivariate clusters and outliers of spatial association

Daniele Oxoli[1] | Soheil Sabri[2] | Abbas Rajabifard[2] | Maria A. Brovelli[1]

[1]Department of Civil and Environmental Engineering - Geomatics and Earth Observation laboratory (GEOlab), Politecnico di Milano, P.zza Leonardo da Vinci 32, Milano, 20133, Italy

[2]Department of Infrastructure Engineering - Centre for SDIs and Land Administration (CSDILA), The University of Melbourne, Melbourne, VIC, 3010, Australia

Correspondence
Daniele Oxoli, Department of Civil and Environmental Engineering - Geomatics and Earth Observation laboratory (GEOlab), Politecnico di Milano, P.zza Leonardo da Vinci 32, Milano, 20133, Italy
Email: daniele.oxoli@polimi.it

Funding information
URBAN-GEO BIG DATA. PRIN Project - Italian Ministry of Education, University and Research (MIUR)— ID. 20159CNLW8.

The detection of spatial cluster and outliers is critical to a number of spatial data analysis techniques. Many techniques embed spatial clustering components with the aim of exploring spatial variability and patterns in a dataset, caused by the spatial association that generally affects most of the spatial data. A frontier challenge in spatial data analysis is to extend techniques - originally designed for univariate analysis - to a multivariate context, with the intent of coping with the increasing complexity and variety of modern spatial data. This paper proposes an exploratory procedure to detect and classify clusters and outliers in a multivariate spatial dataset. Cluster and outliers detection relies on lately introduced multivariate extensions of the well-established Local Indicators of Spatial Association statistics. Two new indicators are proposed enabling respectively the classification of multivariate clusters and outliers, that was not directly achievable with any established technique. The procedure is fully implemented using Free and Open Source Geospatial Software and libraries. The raw source code is made available for future reviews and replications. Empirical results from early applications on both synthetic and real spatial data are discussed. Advantage and limitations of the introduced procedure are outlined according to the empirical results.

KEYWORDS
Spatial Association, Spatial Clustering, Multivariate Spatial Statistics, Geographic Information Systems, Free and Open Source Software

1 | INTRODUCTION

The identification of significant clusters in a dataset is key to many scientific fields connected to data analysis (Grubesic *et al.*, 2014). Indeed, the increasing complexity and richness of modern data have enforced the need for tools and methods to facilitate this task (Hesse *et al.*, 2015). Cluster analysis is central to manifold disciplines including statistics (Kaufman and Rousseeuw, 2009; Anderberg, 2014), ecology (Legendre and Fortin, 1989; Wiegand and Moloney, 2013), and geography (Anselin and Getis, 1992; Anselin, 1995) among others. The popularity of cluster analysis is mostly connected to data grouping options that can yield readily and valuable insights into underlying structures in the data (Kettenring, 2006); providing rooms for unexpected discoveries or additional hypotheses generation (Tukey, 1977). Cluster analysis is generally employed with two main purposes (Xu and Wunsch, 2008). The first one is connected to data preprocessing, where clustering is used as an unsupervised procedure for defining data structure and improving the efficiency of prediction or classification algorithms (Grubesic *et al.*, 2014). The second is exploratory, in which clustering is intended for pointing out patterns and motivating new hypotheses on data according to these discoveries (Tukey, 1977).

The ultimate objective of cluster analysis is to fit each observation in a dataset into a specific group or class. These classes are conceived to portray some underlying structure or process associated with the information carried by the observations (Kettenring, 2006). Therefore, fundamental to clustering is a measure of similarity (or dissimilarity) of the observations being grouped. Focusing on Geographic Information (GI), clustering techniques must take into account also the spatial distribution of observations over the study region (Scrucca *et al.*, 2005). Together with the variable measure provided by each observation, the location is used to define the clustering criterion. This is often expressed as a function of the observation values similarity and their relative position in the geographical space (Han *et al.*, 2001). Among the spatial clustering techniques available in the literature (Han *et al.*, 2009), critical to this work are those based on spatial association properties (Grubesic *et al.*, 2014). The spatial association is broadly defined as the measure of the degree at which similar things are also similarly arranged in space (Tobler, 1970). This property characterizes most of the spatial data and produces peculiar correlation effects within variables across the geographical space (Getis, 2008). The spatial association can be measured by means of statistical indicators (Schabenberger and Gotway, 2017) and therefore adopted as clustering criterion. In general terms, the spatial association can be determined with a single measure for a dataset that describes its global spatial association properties. In the case of clustering, local measures have to be considered to define the spatial association properties at each location in the dataset. According to the above, the statistical indicators considered in this work mostly derive from the well-known Local Indicators of Spatial Association (LISA) (Anselin, 1995). LISA are a family of local statistics aiming at testing for the spatial association and identifying local clusters and outliers in a spatial dataset (Lloyd, 2010). The critical role of LISA in spatial data exploration and analysis has been remarked by many authors through a number of applications across diverse research fields. These includes econometrics (Rey, 2001), landscape genomics (Stucki *et al.*, 2017), ecology (Barrell and Grant, 2013) and remote sensing (Wulder and Boots, 1998), among others. To date, most of LISA techniques focus on univariate analysis that is on the analysis of a single spatial variable. Recent researches have extended the spatial association-based clustering concept to multiple variables, hence to a multivariate context (Anselin, 2019). The development of local and multivariate methods for spatial data analysis is a critical aspect of the modern geography and cognate disciplines. According to Fotheringham (1997) and Lloyd (2010), the increasing volume and complexity of the modern GI - nowadays potentially available at any location on Earth - has led to a greater awareness that the univariate analyses are of limited application in the practice; whereas there is a need to investigate local variations in more complex interaction scenarios. Furthermore, the computational and graphical capabilities of modern Geographic Information Systems (GIS) software have uncovered terrific opportunities to deal with large data volumes as well as to produce local and *"mappable"* statistics that are generally computationally demanding than well-established summarizing or global methods (Efron and Hastie, 2016). Nevertheless, the establishment of multivariate methods for spatial data analysis and mapping raises a number of issues connected to the objective replicability of experiments - that often require analysts to introduce multiple assumptions based on their experience - and to the critical interpretation of multivariate patterns that might not be trivial for non-specialist users (Dray *et al.*, 2011).

In view of the above, the objective of the current paper is to investigate recent advances in local multivariate spatial association-based clustering techniques and to suggest complementary methods for multivariate spatial clusters and outliers classification and mapping. The main purpose of suggesting these complementary methods is to enable the possibility of automatically portray features of a multivariate spatial pattern, such as the numerical characteristics of its underlying observations, by providing analysts with an additional exploratory tool to better understand both data and their spatial interactions. In principles, the proposed methods can be embedded in manifold spatial data analyses and GIS-based applications requiring multivariate patterns being objectively highlighted or described as a preparatory step for statistical processing or mapping of spatial processes that involve multiple variables. The methods are spelt out in their mathematical formulation and tested on both synthetic and real spatial data. Synthetic data consists of a vector grid including observations drawn from four different continuous variables, of which spatial arrangement simulates artificial spatial association features. Real data consists of a polygons vector layer describing census areas for the Metropolitan Melbourne (Victoria, Australia) containing observations of three socio-economical indexes connected to the social vulnerability. The GIS technologies used to carry out this work are selected by exclusively considering cutting-edge Free and Open Source Software (FOSS) and libraries. The role of FOSS within the current work is outlined in connection with the benefits deriving from their systematic use into the scientific research (Brovelli *et al.*, 2017); in particular for promoting fair access and enabling the revision of the experimental procedure here proposed. According, the raw Python code developed for the experiments is made available through the GitHub (https://github.com/opengeolab/mvar_spatial_association_mapping) by providing the technique with the possibility of being improved and replicated. The outcomes are critically discussed alongside possible applications and future directions of the local multivariate spatial association-based clustering techniques.

Focusing on the key points mentioned above, the rest of the paper is organized as follows. Section 2 includes a wider overview of the spatial association concept in both traditional and emerging spatial data analysis; with a focus on spatial clustering. In Section 3, the considered methods are described mathematically. A description of the software tools relevant for this work is included in Section 4. An application of the methods on synthetic data is reported in Section 5, while Section 6 reports an experiment on real data. Section 7 includes key conclusions and discussion on the future directions of this work.

2 | BACKGROUND

The analysis of spatial association - also known as spatial autocorrelation in the literature - has challenged geographers and spatial statisticians over the last 70 years. Since its applications vary considerably from field to field, many analysis methods have been created with different purposes. 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. Goodchild (1986); Anselin and Griffith (1988); Dubin (1998). A closer definition to the context of GIS science is derived from Getis (2010) 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 one or more variables"*. This relationship characterises most of the spatial observations that are measured as a result of nonstationary spatial processes, which take place in the real geographical space (Ord, 2004). Indeed, there are many instances in which the location of an observation affects its behaviour. Snippets of evidence of the spatial association can be intuitively disclosed through simple examples. Housing prices are one among them (Tse, 2002). 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 its a-spatial features such as the building type. On one hand, this behaviour can be seen as a nuisance as it complicates traditional statistical tests by violating the independence assumption for the observations (Dubin, 1998). On the other hand, once the structure of the spatial association is estimated, it can be embedded into any prediction technique, such as the Kriging (Cressie, 1990), thereby improving its accuracy. A fuller treatment of this topic can be found e.g. in Cliff and Ord (1972) and Legendre (1993).

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 to asses both the presence and the degree of the spatial association in a spatial datasets as well as to integrate corrections for the spatial association into spatial modelling (Dormann, 2007). The assumption behind global methods is that spatial association properties are the same across the region of interest. This deficiency may mask spatial variations in the data therefore preventing analysts to detect inner pockets of spatial instability. To that end, 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 (Getis, 2010). Moreover, the growing availability of spatial data at a finer resolution as well as covering large areas, such as high-resolution satellite imagery, continent-wise road networks, etc. has uncovered the need for local methods (Lloyd, 2010). This is because of the probability that regions with different properties would be encountered or considered within the modern spatial analysis has inevitably increased. An example of a popular local method for quantitative spatial data analysis - embedding some of the spatial association concepts - is the Geographical Weighted Regression (GWR) (Brunsdon *et al.*, 1996) which allows regression parameters varying in space.

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 spatial data analysis, and in particular to exploratory techniques such as the spatial clustering. Generally speaking, a spatial cluster can be defined as a geographically bounded group of spatialized observations of sufficient size and concentration that is unlikely to have occurred by chance. Central to the topic is - once again - the need for exploring the local spatial association affecting the data. If the variables of interest (e.g. precipitations, population density, traffic jam, etc.) show significant different spatial association properties along the study region, observations can be grouped according to these properties that is the rationale behind any spatial association-based clustering technique. Restricting our attention on continuous numerical variables, best-known measures of local spatial association are connected to a small family of statistics called LISA. Traditional LISA are the local Moran's I and the local Geary's c statistics. These share a common general formulation (see Section 3.1) and provide measures of spatial association that allow evaluating the existence of spatial clusters in a dataset. Originally, these statistics were designed as global measures of spatial association (Moran, 1950; Geary, 1954). Nevertheless, their local versions have been proposed starting from the 90's (Anselin, 1995). More recently, an extension of LISA to cope with multivariate spatial association analysis has been proposed by Anselin (2019) which consists of a modified version of the local Geary's c statistic, useful to evaluate the existence of spatial patterns in a multivariate dataset. In Section 3.2, the formulation of the multivariate local Geary's c is included. For the sake of completeness, additional local statistics of spatial association have been proposed in the literature. Most popular ones are e.g. the Getis-Ord G (Ord and Getis, 1995) for continuous numerical spatial variables, the local Join Count (Anselin, 1995; Boots, 2003) designed for binary spatial variables, and the geographically weighted Colocation Quotient (Cromley *et al.*, 2014) conceived for categorical spatial variables. The latter statistics are of marginal interest for the current work and therefore they will not be further considered in this paper.

In general terms, the LISA summarize spatial similarity - i.e. contiguity - and numeric similarity of observations into a single statistic of which significance is traditionally inferred against the hypothesis of Complete Spatial Randomness (CSR) by means of conditional random permutations (Besag and Diggle, 1977) (see Section 3.1). Both univariate and multivariate LISA computations allow only detecting whether positive or negative spatial association affects each observation in the dataset. Positive spatial association stands for the significant presence of similar observations in the neighbourhood of the considered observation location, therefore a high probability for the considered observation to belong to a spatial cluster. Conversely, negative spatial association implies a significant presence of dissimilar observations close in space, thus the likely the considered observation being labelled as a spatial outlier. Considering spatial clusters, the significant similarity detected by the LISA can be either due to the presence of low or high values close in the geographical space. The definition of low and high values refers always to the mean of the distribution of the analysis variable that is observed in the region of interest (Anselin, 1995). With this in mind, clusters can be further classified into two classes which are namely clusters of high values and clusters of low values. A similar classification can be applied to spatial outliers. Indeed, the significant dissimilarity detected by the LISA can be due to the presence of a low value for the considered observation close in space to higher values by obtaining thereof a low-high outlier. Vice-versa, a high-low outlier is defined. As stated in the above, this classification for both clusters and outliers cannot be directly achieved by means of LISA computations. Traditionally, the classification is performed using linked techniques such as the Moran scatterplot (Anselin, 1996, 1999) applied alongside the local Moran's I computations. The same technique can be combined also to the local Geary's c by means of brushing and linking operations on the identified clusters and outliers with the Moran scatterplot, as suggested by Anselin (2019). Additional information on the clusters and outliers classification by using the Moran scatterplot is included in Section 3.1.

The main limitation of the Moran scatterplot is that no multivariate clusters and outliers classification can be achieved; as highlighted later in Section 3.2. The most common approach in multivariate spatial association analysis is to introduce a preliminary dimensionality reduction step - such as the Principal Component Analysis (PCA) - and then apply univariate or bivariate techniques on the reduced set of spatial variables (Dray *et al.*, 2011). To overcome the above limitations, this paper presents an automatic procedure to classify multivariate clusters and outliers, detected by means of the multivariate local Geary's c (see Section 3.2). The procedure is based on comparisons between centrality measures (i.e. the median and the mean) for the variable distributions within clusters and outliers locations. The procedure for classifying and mapping multivariate spatial clusters and outliers is spelt out in Section 3.3. In overall terms, the aim of this work is to empower any data exploratory experiment that requires a number of continuous spatial variables being considered simultaneously without any prior manipulation, thus favouring the analysis and display of their underlying covariate patterns.

3 | METHODS

This section is dedicated to the mathematical formulation of LISA by including also recent developments and extensions of these methods to a multivariate context. Limitations of traditional techniques for multivariate clusters and outliers classification are also discussed and an automatic procedure to perform such is proposed.

3.1 | Traditional LISA

Among LISA, the most popular is the local Moran's I (Equation 1) Anselin (1995), which is a local version of the Moran's Index (Moran, 1950).

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

Where n is the number of locations in the dataset and z_i and z_j are the standardized observation values, such that their mean is zero and their variance is one, at two locations i and j. $W_{i,j}$ is the so-called spatial weights matrix whose values define whether location j is a geographic neighbour - or it is not - of location i. This is often expressed by a symmetric matrix having values, e.g. ones, at i (i.e. row) and j (i.e. column) position if i and j locations are defined as neighbours, and zeros elsewhere. However, a number of neighbouring or contiguity relationships can be adopted by the analyst through rules such as a threshold distance for point data or edges and corners congruence for areal data. A fuller description of geographical weighting schemes can be found e.g. in Getis and Aldstadt (2004).

A significant and positive value for I_i 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. Inference on the CSR hypothesis is provided through conditional random permutations tests by assigning a pseudo p-value (p_i) to each location. 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 $n - 1$ locations, and recompute the local spatial association statistic. By repeating m times this process, an empirical reference distribution for the statistic under the CSR hypothesis is obtained at each location. Using this reference distribution, the p_i of the statistic at location i can be computed as in Equation 2 (Phipson and Smyth, 2010).

$$p_i = \frac{b+1}{m+1} \tag{2}$$

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. The smaller the pseudo p-value the stronger the rejection of the CSR hypothesis. In turn, this means a higher probability that location i 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 (Anselin, 2019) because it is not likely to properly reflect the actual Type I error, that is the case when the null hypothesis (i.e. CSR) is true but it is rejected thus a false positive is encountered (Efron and Hastie, 2016). Due to the computational procedure adopted in the conditional permutations approach, many of the values used to simulate a local measure of spatial association at each location i, are used again to test on a neighbouring location by producing a large number of correlated tests. Ironically, this means that by searching for evidence of spatial association the tests are affected by the spatial association themselves (Getis, 2010). 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 to control the error rate that is known as the False Discovery Rate (FDR) (Boots, 2002). One of the most popular is the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). 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 FDR 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 $p_{i=1} \leq p_{i=2} \leq ... \leq p_{i=n}$. 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 is verified, where n is the number of observations. All observations with i lower or equal than i_{max} are then considered significant for rejecting the null hypothesis.

$$p_{i,significant} \leq (\frac{i_{max}}{n})\alpha \tag{3}$$

Once significant clusters and outliers are detected by means of the local Moran's I, these can be classified into high and low values clusters or low-high and high-low outliers, according to what anticipated in Section 2. The classification is performed by plotting on a Cartesian plane each couple $(z_i; W_{i,j} z_j)$. The resulting plot is known as the Moran scatterplot. Basically, this couple is composed of the standardized observation values at the location i and its spatial lag (Anselin, 1996). This implies classifying local clusters and outliers base on the linear correlation among the two components. Depending on the quadrant of the plane at which the point representing the above couple is found, and belonging this couple to a significant cluster or outlier location, the classification is carried out as reported in Figure 1. The classification into quadrants is used to display results on a map by assign a proper visualization style to the spatial dataset using e.g. a GIS software.

The local Moran's I is suitable for both univariate and bivariate analyses. In the bivariate setting, the spatial dataset has two variables observed at each location. In Equation 1, the z_j is substituted with the standardized values of the second variable. Clearly, the standardization procedure becomes here necessary in order to scale and compare the different variables. By doing so, a significant and positive value for I_i indicates that location i has the value of the first variable similar with values of the second variable in its j neighbours. Location i is - therefore - a joint spatial cluster for the two variables. A significant and negative value for I_i hence location i being a joint spatial outlier for the two variables. Clusters and outliers classification can be achieved by means of the Moran scatterplot as for the univariate case.

The second LISA outlined in Anselin (1995) is the local Geary's c. Its formulation derives from the decomposition of the global parent statistic, the Geary's c, proposed in the '50s by Geary (1954). While the local Moran's I provides a local measure of spatial association as cross-products among a focal location and its neighbours, the local Geary's c expresses the spatial association as a weighted average of squared distances in the observation space between values at each location i and those at its neighbouring locations j, as shown in Equation 4.

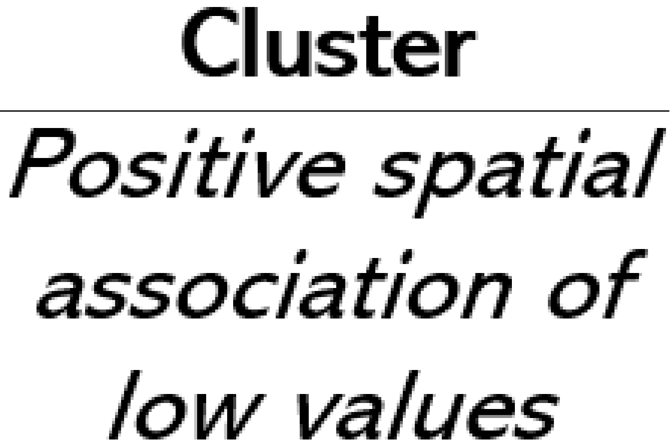


FIGURE 1 Clusters and outliers classification through the Moran scatterplot.

$$c_i = \sum_{j=1}^{n} W_{i,j}(z_i - z_j)^2 \;; i \neq j \tag{4}$$

Where z_i and z_j are the standardized observation 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 statistic focuses on dissimilarity rather than correlation (Anselin, 1995). Therefore, a significant c_i value lower than its expected value indicates similarity or positive spatial association. On the other way round, a significant c_i value larger than its expected value indicates dissimilarity or negative spatial association. The expected value of c_i can be either computed analytically, e.g. the sample mean, or from an empirical reference distribution such as the one derived by the conditional permutations used for significance testing. Finally, clusters and outliers classification can be indirectly achieved by means of brushing and linking operations on the significant locations detected by means of c_i to the Moran scatterplot (Anselin, 2019).

3.2 | Multivariate LISA

In a multivariate context, i.e. when k variables are observed at each location and their spatial association needs to be investigated simultaneously, no consolidated methods have been made available in the literature. Recently, an extension of LISA has been proposed by Anselin (2019), enabling multivariate spatial association analyses. This is an extension of the local Geary's c statistic. The multivariate extension of the local Geary's c collapses the k squared distances between the observation values at location i and that at each neighbouring location j into a weighted sum by providing a single value for the statistic at each location of a spatial dataset, as shown in Equation 5.

$$c_{k,i} = \sum_{v=1}^{k}\sum_{j=1}^{n} W_{i,j} d^2_{v_{i,j}} \;;\; d^2_{v_{i,j}} = (z_{1,i} - z_{1,j})^2 + \ldots + (z_{k,i} - z_{k,j})^2;\; i \neq j \tag{5}$$

Where $d^2_{v_{i,j}}$ is the k-dimensional squared distance in the observations space between the standardize z_v observation values at locations i and j, and $W_{i,j}$ is the spatial weights matrix. The interpretation of spatial association using the multivariate local Geary's c follows the one explained for the univariate case. A significant value of $c_{k,i}$ that is lower than its expected value suggests positive spatial association, whereas a significant and higher value suggests negative spatial association. The expected value can be either computed analytically or from an empirical reference distribution derived by the same conditional permutations experiment used for significance testing, as for the univariate case explained before.

Clusters and outliers classification, such as the one introduced with the Moran scatterplot, is not achievable in this case, leading to a strong limitation in exploring the spatial association by the exclusive use of the multivariate local Geary's c. This because the Moran scatterplot is a bi-dimensional plot that allows comparing up to two variables per time, thus preventing its adoption in the context of multivariate analysis. For this reason, the main objective of the current work is to suggest a classification technique for spatial clusters and outliers that overcomes the limitation of the Moran scatterplot in multivariate analyses.

3.3 | Multivariate clusters and outliers classification

The proposed technique for multivariate clusters and outliers classification is based on a comparison between centrality measures (i.e. the median and the mean) for the variable distributions within cluster and outlier locations. Two indicators are proposed for classifying clusters and outliers that we called respectively Mm_c (Median-mean indicator for clusters) and Mm_o (Median-mean indicator for outliers). The proposed technique is conceived and outlined by considering a generic multivariate tabular dataset where each row represents a single spatial location and each column contains observations drawn from one spatial variable. This general data model is common - or it can be commuted - to many standard spatial data formats currently in use. The proposed technique is intended to be applied on the cluster and outlier locations that result from the computation of the multivariate local Geary's c. The final output consists of a classification for multivariate spatial clusters and outliers that inherits the concept of the Moran scatterplot quadrants for the univariate case, while allowing mapping local multivariate spatial association patterns.

The Mm_c indicator (Equation 6), enabling the classification of multivariate clusters is defined as follows. We call X the $n \times k$ matrix containing the standardized observations of the k original variables at each of the n locations of the dataset. For each location i that resulted in a cluster from the multivariate local Geary's c computations, a $r \times k$ subset $X_{i,j}$ containing the observations at location i and at its geographic neighbours j (spatial lag of i) is extracted. The idea is to compare the mean of the vector of the column-wise medians of $X_{i,j}$ ($\mu_{X^M_{i,j}}$) with the mean of the vector of the column-wise medians of X (μ_{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 to account for possible skewed distribution of the analysis variables. This reduces the bias in estimating the central tendency values for the comparison, due to the influence of possible outliers (Von Hippel, 2005). The proposed local indicator $Mm_{c,i}$ consists of a simple difference between $\mu_{X^M_{i,j}}$ and μ_{X^M} at each cluster location i (Equation 6). A positive value of the $Mm_{c,i}$ depicts a higher local mean of the medians than the one of the whole study region, hence a cluster of high values. A negative value of the $Mm_{c,i}$ depicts a lower local mean of the medians than the one of the whole study region, hence a cluster of low values.

$$Mm_{c,i} = \mu_{X^M_{i,j}} - \mu_{X^M}, \rightarrow \begin{cases} Mm_{c,i} > 0, & \text{High values cluster} \\ Mm_{c,i} \leq 0, & \text{Low values cluster} \end{cases} \tag{6}$$

The Mm_o indicator (Equation 7), enabling the classification of multivariate outliers, is defined as follows. We call X the $n \times k$ matrix containing the standardized observations of the k original variables at each of the n locations. For each location i that resulted in a outlier from the multivariate local Geary's c computations, two subsets X_i and X_j, containing respectively the $1 \times k$ observations at location i, and the $(r-1) \times k$ observations at its geographic neighbours j (spatial lag of i), are extracted. In this second case, the idea is to compare the mean of the vector X_i (μ_{X_i}) with the mean of the vector of the column-wise medians of X_j ($\mu_{X^M_j}$). This implies comparing the mean of the observations at an outlier location with the mean of the medians of its neighbouring locations only. The median is again suggested for the same reasons explained before. The proposed local indicator $Mm_{o,i}$ consists of a simple difference between μ_{X_i} and $\mu_{X^M_j}$ at each outlier location i (Equation 7). A positive value of the $Mm_{o,i}$ depicts a higher local mean than the mean of the medians from the neighbouring locations, hence a possible high-low outlier. A negative value of the $Mm_{o,i}$ depicts a lower local mean than the mean of the medians from the neighbouring locations, hence a possible low-high outlier.

$$Mm_{o,i} = \mu_{X_i} - \mu_{X^M_j}, \rightarrow \begin{cases} Mm_{o,i} > 0, & \text{High-low outlier} \\ Mm_{o,i} \leq 0, & \text{Low-high outlier} \end{cases} \tag{7}$$

Results form the Mm_c and the Mm_o computations can be used to enrich clusters and outliers maps obtained from the multivariate local Geary's c. With this additional information, multivariate clusters and outliers can be classified thus producing a multivariate LISA map comparable to that obtained from the combination of local Moran's I with the Moran scatterplot in the univariate setting.

Despite the number of steps required by the proposed procedure, perhaps comparable to those required by the Moran scatterplot, the definition of an automatic strategy for enriching multivariate spatial association maps by means of clusters and outliers classification is achieved. It's important to remind that the Moran scatterplot produces a classification based on the linear correlation properties of a variable and its spatial lag at a specific location. The Mm indicator relies instead on comparisons between centrality measures of the variable distributions at a specific location and its spatial lag with respect to the global distribution (clusters) or the variable distributions at a specific location with respect to the one in its spatial lag (outliers). In a multivariate context, centrality measure may mask trade-off among variables and therefore results have to be intended as an exploratory tool rather than in a rigorous statistical sense. The same applies to the multivariate LISA technique on which the whole procedure is based, i.e. the multivariate local Geary's c, as argued also in Anselin (2019).

4 | SOFTWARE TOOLS

Any data exploratory technique has a strong connection with graphical and interactive software tools (Keim *et al.*, 2006). In the case of spatial data, GIS software is naturally entitled as the best tool to perform data analysis, interaction, and visualization. Focusing on traditional LISA, many software implementations are available. These include libraries for popular programming languages, stand-alone statistical software, and GIS software modules. In the following, cutting-edge solutions dedicated to LISA mapping are pointed out by exclusively considering Free and Open Source Software (FOSS). The underlying reason behind this choice is due to the fair distribution policy of FOSS. Indeed, this kind of software and libraries are released with open licenses (https://opensource.org/licenses) - in contrast to proprietary software - thus allowing users free access to thereof functionalities and the source code. This provides analysts with plenty of opportunities to investigate, test, and extend spatial analysis methods, therefore representing a meaningful asset for this work in which an attempt to introduce new data processes is undertaken. Indeed, code has to be considered as a critical research object in any spatial technique development work (Ertz *et al.*, 2014). The possibility of using, modifying, and exposing to the scientific community the source code developed to test the proposed classification technique, provides the analysis with the possibility of being replicated, empowered, and improved.

Considering the most popular geospatial FOSS libraries, LISA functionalities are provided by the spdep library (https://cran.r-project.org/web/packages/spdep) for the R programming language, and by the PySAL library (https://pysal.readthedocs.io) for Python. Among stand-alone statistical FOSS platforms, GeoDa (https://geodacenter.github.io) provides with plenty of LISA as well as spatial data exploration functionalities and supports most of the standard spatial data formats. Concerning FOSS GIS, like the popular QGIS (https://qgis.org), an experimental LISA mapping plugin, developed by some of the authors and called Hotspot Analysis (Oxoli *et al.*, 2017), is available (https://plugins.qgis.org/plugins/HotspotAnalysis).

To the extent of the authors' knowledge, multivariate LISA functionalities have been recently made available only into the GeoDa software. No other solutions dedicated to multivariate LISA exist, also by considering proprietary software. Unfortunately, GeoDa provides with less support for result files post-processing than the other mentioned FOSS software and libraries. Therefore, to test the proposed clusters and outliers classification technique, a custom Python script was developed which implements both the multivariate local Geary's c and the Mm computations (see sections 5 and 6). The script allows applying the full procedure for multivariate cluster and outliers detection and classification on a vector spatial layer in Shapefile format. The script takes advantage of the GeoPandas library (http://geopandas.org) for generic data operations such as read, write and metadata access. The script includes custom functions based on the PySAL routines, coupled with statistical and numerical operations based on the NumPy (https://numpy.org), the SciPy (https://scipy.org) and the statsmodel (https://www.statsmodels.org) libraries. Both the code and test data are available on GitHub (https://github.com/opengeolab/mvar_spatial_association_mapping).

5 | EXPERIMENT ON SYNTHETIC MULTIVARIATE SPATIAL DATA

Testing for the effectiveness of the proposed technique necessitates the use of synthetic spatial data that contains known and established local spatial association characteristics. To that end, the approach used to generate a dataset with known spatial clusters is reported in the following. Both the multivariate local Geary's c and the Mm are then computed on this dataset and the experimental results are compared to the expected outcomes.

5.1 | Synthetic data generation

The generation of synthetic data for testing the proposed technique was achieved as follows. A regular spatial grid with 100x100 pixels in Shapefile format was created. For this experiment, four spatial variables were considered. Observations from each variable at each pixel were created by means of sampling from four continuous standard normal distributions generated with different skewness. This artefact allowed to account for spatial variables showing different distribution characteristics that are likely to be encountered in the real multivariate spatial analysis. Both distributions generation and sampling were performed by means of SciPy (Figure 2a). The spatial grid was arbitrarily divided into eight sub-regions to simulate spatial association (Figure 2b). Namely, observations were assigned to the pixels of each sub-region by selecting them according to their belonging quartile interval (Figure 2b). Hence, the similarity between observations to simulate spatial association was here modelled using, as a rule, the membership of observations to the same range of the standardized variable distributions. These ranges are - in this case - the four distribution intervals bounded by the quartiles. The spatial similarity was instead obtained by simply bounding neighbouring pixels into the selected sub-regions. This allowed defining homogeneous areas in terms of observation values that result in known and overlapping spatial clusters for all the variables. Summary statistics of both variables and their quartile intervals are reported in Table 1.

TABLE 1 Summary statistics of the simulated variables (A,B,C and D) used for for generating the synthetic dataset, including quartiles (Q1, Q2 and Q3), and quartile interval lengths (L) . The variables are standardized, such that their mean is zero and their variance is one. Measures are provided as deviation from the mean.

	A	B	C	D
Skewness	0,607	0,255	-0,096	-0,762
Min	-2,199	-3,499	-3,395	-4,564
Max	4,556	3,210	3,461	2,239
Q1	-0,669	-0,552	-0,583	-0,413
Q2	-0,124	0,0696	0,033	0,167
Q3	0,589	0,659	0,663	0,683
L[Min,Q1]	1,530	2,947	2,812	4,151
L[Q1,Q2]	0,545	0,622	0,616	0,580
L[Q2,Q3]	0,713	0,589	0,630	0,516
L[Q3,Max]	3,967	2,551	2,798	1,556

According to the assigned quartile interval, known clusters of either high or low values could be simulated. By definition, the Mm indicator for clusters classification (Mm_c) compares centrality measures (i.e. means and medians) at each location in the dataset and its geographic neighbours to the ones of the whole study region. Therefore, the aggregation in a sub-region of standardized observations belonging to the same quartile interval aimed at controlling the type of clusters that are expected in each sub-region. In particular, clusters are expected within each sub-region due to the numeric similarity of the observations there assigned. High values clusters are mainly expected into sub-regions where observations from the 3rd and 4th quartile intervals are present. This because a concentration of observation values higher than the mean of each variable was generated. On the other way round, low values clusters are mainly expected into sub-regions where observations from the 1st and 2nd quartile intervals are present.

Concerning spatial outliers, these cannot be simulated by applying the same logic used for clusters. Indeed, negative spatial association and - in turn - a spatial outlier is found where an observation shows significantly different values with respect to its neighbouring values, even though it may not be significantly different from the entire population (Lu *et al.*, 2003). Thereby, spatial outliers may show up due to a number of observation arrangements in space which cannot be coded into simple rules such as for spatial clusters. Indeed, the Mm indicator for outliers classification (Mm_o) compares centrality measures (i.e. means and medians) at each location in the dataset to the ones of its geographic neighbours and not to the ones of the whole study region such as for clusters. In view of the above, outliers are expected mainly along sub-regions borders but also within the sub-regions. The latter may correspond to the case where extreme observations from the same quartile interval of differently distributed variables were assigned by chance to neighbouring pixels during the generation of cluster areas. More details are reported in the following section.

Random observations drawn from each variable were assigned instead to the pixels located outside the eight sub-regions (Figure 2b). This means that also spatial randomness was simulated at known locations of the spatial grid.

A multivariate spatial dataset was thus obtained and used to test the capability of both the multivariate local Geary's c and the Mm indicator to respectively detect and classify multivariate clusters and outliers of spatial association. The synthetic dataset is available on GitHub (https://github.com/opengeolab/mvar_spatial_association_mapping/tree/master/sample_data). The experiment on the synthetic dataset was performed by selecting a 1st order edges and corners contiguity rule - known as queen contiguity - for generating the spatial weights matrix. Hence, each pixel has eight neighbour pixels on a regular grid, except for the ones placed on the grid perimeter. Row-standardized weights were adopted (Getis and Aldstadt, 2004). The spatial weights matrix was created by exploiting the dedicated PySAL functionalities. The CSR hypothesis was tested by means of 99999 permutations with a selected significance level of 0.0001. Pseudo p-values were corrected by applying the Benjamini-Hochberg FDR procedure using the dedicated functionality of the statsmodels library. The resulting multivariate clusters and outliers map is included in Figure 3.

5.2 | Results discussion

By observing the map of the results in Figure 3, it can be argued that clusters are located within each of the eight sub-regions. High values clusters emerge exclusively from the sub-regions where observations from the 3rd and 4th quartile intervals were associated. In the same way, low values clusters are detected exclusively in the sub-regions that include observations from the 1st and 2nd quartile intervals. The results for clusters are perfectly aligned to what expected from the synthetic dataset. The above provides evidence of the Mm_c indicator reliability for multivariate clusters classification.

Concerning pixels where spatial randomness was simulated, the local multivariate Geary's c provided with robust local statistics for CSR hypothesis testing. Indeed, all those pixels resulted not significant according to the significant level adopted in this experiment (see Section 5.1).

A more complex situation is found at the borders between the sub-regions. Specifically, sub-regions 1 and 2 (Figure 3) include observations from the 3rd and the 4th quartile intervals and therefore high values clusters. However, not significant pixels are also present along the border between these two sub-regions. The same happens between sub-regions 7 and 8. This may be due to the variability that is artificially introduced into the border pixels with respect to their neighbours by assigning observations from different quartile intervals, even tough contiguous quartile intervals were considered. This variability may result in higher square distances between observations and - in turn - in not significant values of the local multivariate Geary's c. Between sub-regions 3 and 4, where observations from the 4th and 1st quartile intervals respectively are located, the border effect is enhanced due to the higher variability that was introduced into the border pixels with respect to their neighbours than the previous case.

The variability between sub-regions 5 and 6, where observations from the central quartile intervals are located (i.e. 2nd and 3rd quartile intervals respectively), seems to not influence the detection of the clusters. The latter may be due to the presence of similar observations with a small spanning around the mean from each variable. This may be connected to the smaller quartile interval lengths than the 1st and 4th quartile intervals (see Table 1) that may results in a smaller variability in terms of square distances between observations. In the current experiment this similarity generated significant values of the local multivariate Geary's c also along the sub-region boarders. Nonetheless, the Mm_c indicator provided with proper results also in this particular arrangement of the observations. Finally, a small number of not significant pixels is found also within the inner area of the sub-regions. Actually, this happened only into the sub-region 3 (Figure 3) in the current experiment. Once again, this may be connected to the presence of observations from the 4th quartile intervals - affected by higher variability than e.g. the 2nd and 3rd quartile intervals - that were assigned by chance to neighbouring pixels during the generation of this clusters area.

The negative spatial association - i.e. outliers - has been left outside from the result discussion so far. Indeed, results from multivariate spatial outliers require a separate review. As pointed out in the previous section, spatial outliers show up due to a number of observation arrangements in space which cannot be controlled by simple rules such as for spatial clusters. As expected, outliers are found along both inner and outer borders of the sub-regions. Nevertheless, a number of outliers are also scattered within the inner area of the sub-regions. Once again, these outliers may be generated by the variability of observations from the same quartile interval of the different variables, that were assigned by chance to neighboring pixels during the generation of the cluster areas.

Considering the inner area of the sub-regions, high-low outliers are found exclusively within sub-regions 2 and 3 that include observations from the 4th quartile intervals. Low-high outliers are found instead exclusively within the inner area of sub-regions 4 and 8 that include observations from the 1st quartile intervals. Once again, the triggering factor of this spatial effect may be associated to the high variability of observations belonging to those quartile intervals, that is connected to their higher lengths than the 2nd and 3rd quartile intervals (see Table 1). Nevertheless, it can be argued that the Mm_o indicator produced coherent results also for outliers classification. Indeed, the 4th quartile intervals and - in turn - sub-regions 2 and 3 include also the tail observations (above the mean) from each variable. As a consequence, high-low outliers are likely to be found in the aforementioned sub-regions and the Mm_o indicator proved to be capable of classifying them accordingly. The same applies to low-high outliers which are mostly found in sub-regions 4 and 8 that include the tail observations (below the mean) from the 1st quartile intervals.

Finally, to observe the susceptibility of the proposed procedure to the spatial weighting schema, additional multivariate clusters and outliers maps were included in Figure 4. These maps were created by using the multivariate local Geary's c and the Mm classification but considering different spatial weighting rules than the one adopted in the previous example. Namely, the 1st order corners contiguity rule - known as rook contiguity - (Figure 4b) and the Kernel Gaussian weighting rule (Figure 4c), using a bandwidth equal to two times the pixel size of the synthetic dataset grid (Getis and Aldstadt, 2004). These two maps can be compared to the one obtained from the queen contiguity spatial weights matrix (Figure 4a). All maps were obtained by adopting row-standardized weights, 99999 permutations with a significance level of 0.0001 for the CSR inference, and FDR-corrected pseudo p-values. Despite minor differences in the location of clusters and outliers - anyhow foreseen due to different geographic neighbours definitions - expected patterns are preserved in all the maps by suggesting the robustness of the procedure independently to the adoption of different spatial weighting rules.

Despite based on qualitative investigations, the disclosure of results explains most the spatial association features affecting the dataset. However, a general explanation of the patterns cannot be achieved through qualitative considerations. The presence of outliers, as well as unexpected not significant locations, should be investigated locally and case-wise to outline a more rigorous explanation about the observations arrangement and the trade-off among variables which produce such patterns.

6 | CASE STUDY: SOCIAL VULNERABILITY IN MELBOURNE

In this section, an experiment on a real multivariate spatial dataset is presented. Spatial variables considered were three social vulnerability indexes for the Metropolitan Melbourne. To accomplish that, both procedure and software tools already employed in the experiment on the synthetic dataset were here adopted. As stated in sections 1 and 2, spatial association analysis is critical to many disciplines including social science and urban planning, to which extent results form the proposed case study can be operationally applied. However, the purpose here was to investigate performances and limitations of the proposed procedure on real multivariate spatial data rather than produce evidence for a scientific analysis of the social vulnerability for the Metropolitan Melbourne. The dataset characteristics are introduced in the following and the results are discussed according to the above considerations.

6.1 | Case study data and experiment setup

Social vulnerability is a complex phenomenon driven by a multitude of factors. Despite the fact that formal measures of vulnerability are not currently available (Smit and Wandel, 2006), recurrent factors in the scientific literature include broader social, economic and political backgrounds that spatially interact by producing observable patterns (Lee, 2014). The selection of indexes for this case study aimed at conveying some of these different factors in order to purely explore their spatial interaction. No other motivations were behind the indexes selection. Indeed, these were here used as an example of the multiple spatial variables potentially contributing to a single spatial process that the proposed analysis aims at uncovering and describing.

The indexes considered for the analysis were: the Vulnerability Assessment for Mortgage, Petrol and Inflation Risks and Expenditure (VAMPIRE), the Index of Relative Socio-economic Disadvantage (IRSD) computed by the the Australian Bureau of Statistics, and the total unemployment rate provided by the Organization for Economic Cooperation and Development (OECD) records. Data refers to the Year 2011. Indexes were aggregated at a census tract level (areal data) corresponding to the Statistical Area Level 2 (SA2) defined by the Australian Statistical Geography Standard (ASGS). Details on the indexes are included in Table 2. Aggregated indexes at the SA2 level were retrieved directly from the data download platforms of the providers. According to the mathematical definition of each index, a preliminary manipulation of the data was required. Namely, the scales of the variables had to be adjusted to meet the analysis requirements. In this case, high values of both the VAMPIRE and the OECD indicate lousy performance in terms of social vulnerability for a census parcel whereas the lower the values, the better the performance. Conversely, low values for the IRSD mean higher social vulnerability performances. Therefore, the inverse of the IRSD observations - computed as one divided by the original value of the observations - is considered for the computations.

The spatial distribution of the indexes is reported in Figure 5a where the colour ramp for the IRSD is inverted to graphically account to what stated above. A row-standardized queen contiguity spatial weights matrix was created. A schematic of the resulting spatial weighting and the neighbours distribution are reported respectively in figures 5b and 5c. Being derived from irregular geometries (i.e. census tracts), the neighbours distribution shows higher variability than e.g. the one resulting from the regular grid used in the previous experiment. Moreover, no prior information on the spatial association affecting the variables was available, conversely to the synthetic dataset. Insights into the underlying spatial patterns of each index were obtained by means of univariate spatial association mapping using the local Moran's I and the Moran scatterplot for clusters and outliers classification. Results were computed using the GeoDa software are reported in Figure 6.

The multivariate spatial association mapping was performed using the custom Python script described in Section 4.

TABLE 2 Social vulnerability indexes used in the multivariate analysis.

Index name	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, https://data.aurin.org.au)
IRSD	The Index of Relative Socio-economic Disadvantage is a general socio-economic index that summarizes 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 score was considered. This to disambiguate its interpretation in terms of social vulnerability performance with respect to the other two indexes. (2011, https://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)

The resulting multivariate clusters and outliers map is included in Figure 7. The CSR hypothesis for both univariate and multivariate experiments was tested by means of 99999 permutations with a selected significance level of 0.0001. Pseudo p-values were corrected by applying the Benjamini-Hochberg FDR procedure. Resulting maps in Figure 6 maintain the colour schema provided by the GeoDa software, which nevertheless is consistent with the one selected for the multivariate mapping (see Figure 7). This allows distinguishing results between univariate and multivariate maps and - at the same time - producing a comparable visualization of patterns.

6.2 | Results discussion

Multivariate clusters and outliers patterns are here qualitative compared with univariate patterns. The comparison aims at investigating the general behaviour of the proposed multivariate spatial association mapping procedure on real data with unknown spatial association properties. This is exactly the purpose of exploratory analysis in which context the proposed procedure is conceived.

Univariate clusters and outliers patterns in Figure 6 reveal a significant spatial association affecting the indexes. According to the definition of the indexes, low values indicate generally better performance than high values in terms of social vulnerability. Hence, low value clusters identify areas wherein vulnerability is expected to be lower than in high value clusters areas, whereas outliers point out locations with significantly different vulnerability conditions than their geographic neighbours. Univariate clusters and outliers maps in Figure 6 depict different spatial association patterns for the three indexes. Assuming that social vulnerability at each census tract is produced by the spatial interaction of more than one factor, the multivariate spatial association mapping was used here to point out areas that are likely to be affected by either high or low vulnerability conditions. To this end, interesting areas affected by different vulnerability conditions were clearly highlighted on the map in Figure 7. The resulting multivariate clusters and outliers map provides with a snapshot of spatial association patterns which are not obvious by observing independently the univariate patterns. The latter statement, derived from this experiment, uncovers the limitations in the use of the proposed procedure. As it was also argued in Anselin (2019), a multivariate pattern is not merely a combination of its underlying univariate counterparts, but it involves complex trade-offs in all the variables dimensions considered for the analysis. Nevertheless, the experiment demonstrates the applicability of the procedure on tradition spatial data providing asset maps that might support further investigations on vulnerability issues that are functional e.g. to social policies evaluation and urban planning. In this specific context, the outcomes may contribute to improving decision-making processes through the generation of prioritization maps for urban design interventions as well as work as a background for further spatial econometric analyses. A deep discussion on the use of the outcomes of this case study is out of scope for the current paper. Additional details on outcomes assessment for the presented case study can be found in Oxoli (2019).

7 | CONCLUSIONS

In this paper, a new procedure for detecting and classifying local multivariate spatial clusters and outliers is presented. The spatial statistics on which the procedure is based - or derived from - are described in details. Two new indicators enabling respectively the classification of multivariate spatial clusters and outliers are proposed. These are intended as complementary tools to multivariate spatial association-based clustering techniques, in particular to the multivariate local Geary's c statistics.

The procedure is tested both on a synthetic dataset - with local spatial association properties artificially simulated thereof - as well as on a real spatial dataset. Empirical results show the capability of the procedure to grasp and highlight multivariate patterns of spatial association. However, results on real spatial data point out limitations in understanding and describing these patterns that may be generated by complex trade-offs among variables in a multivariate setting. These trade-offs cannot be fully investigated by means of the proposed procedure which nevertheless has to be intended as an exploratory tool rather than as a rigorous statistical technique.

Concerning the software side, the source code produced within this work requires a substantial improvement to be integrated into existing computing libraries or GIS software. 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 computational time is advised due to the critical role of this factor to the practical application of the procedure. The same concern applies when considering the analysis of high-resolution multidimensional datasets across large geographic regions; that is one of the frontiers of the modern spatial data analysis. The maturity of FOSS technologies considered in this study - coupled with their scalability, performance and plain interoperability - is promising to accomplish the above challenges.

Extensive testing of the proposed procedure has not been carried out within this work. This leave rooms for further researches that should mainly focus on validating outputs and investigating the analytical implications of multivariate spatial association into the emerging spatial analysis techniques. This revision and improvement phase is encouraged by the accessibility to the source code ensured by the exclusive use of FOSS. Despite its early stage, the procedure 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. Indeed, the possibility of performing multivariate spatial association mapping - through a compact and replicable GIS-based procedure - opens newsworthy opportunities for practitioners of these disciplines, by favouring both the scientific debate on the topic and the embracement of multivariate thinking into all those geographical studies that have exploited only univariate analysis so far.

Applications of multivariate spatial association measures may also complement cutting-edge spatial methods such as object-based image analysis algorithms by favouring an unsupervised modelling of underlying and complex spatial patterns. At the same time, additional applications in the context of data exploration may be investigated, such as on the use of the proposed technique for analysing spatio-temporal patterns. This by considering observations at multiple times of a single spatial variable rather than from multiple variables. All these assumptions will be tackled with priority by the future developments of this work.

acknowledgements

This work was partially supported by URBAN-GEO BIG DATA, a Project of National Interest (PRIN) funded by the Italian Ministry of Education, University and Research (MIUR)— ID. 20159CNLW8.

conflict of interest

The authors declare no conflict of interest.

references

Anderberg, M. R. (2014). *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Vol. 19. : Academic press.

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis*, 27 (2), 93–115.

Anselin, L. (1996). The moran scatterplot as an esda tool to assess local instability in spatial association. In: *Spatial Analytical*. Vol. 4.

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

Anselin, L. (2019). A local indicator of multivariate spatial association: extending geary's c. *Geographical Analysis*, 51 (2), 133–150.

Anselin, L. and Getis, A. (1992). Spatial statistical analysis and geographic information systems. *The Annals of Regional Science*, 26 (1), 19–33.

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

Barrell, J. and Grant, J. (2013). Detecting hot and cold spots in a seagrass landscape using local indicators of spatial association. *Landscape ecology*, 28 (10), 2005–2018.

Benjamini, Y. and 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)*, 57 (1), 289–300.

Besag, J. and Diggle, P. J. (1977). Simple monte carlo tests for spatial pattern. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26 (3), 327–333.

Boots, B. (2002). Local measures of spatial association. *Ecoscience*, 9 (2), 168–176.

Boots, B. (2003). Developing local measures of spatial association for categorical data. *Journal of Geographical Systems*, 5 (2), 139–160.

Brovelli, M. A., Minghini, M., Moreno-Sanchez, R. and Oliveira, R. (2017). Free and open source software for geospatial applications (foss4g) to support future earth. *International journal of digital earth*, 10 (4), 386–404.

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

Cliff, A. and Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical analysis*, 4 (3), 267–284.

Cressie, N. (1990). The origins of kriging. *Mathematical geology*, 22 (3), 239–252.

Cromley, R. G., Hanink, D. M. and Bentley, G. C. (2014). Geographically weighted colocation quotients: specification and application. *The Professional Geographer*, 66 (1), 138–148.

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

Dray, S., Jombart, T. and *et al. (2011). Revisiting guerry's data: introducing spatial constraints in multivariate analysis. The Annals of Applied Statistics, 5 (4), 2278–2299.*

Dubin, R. A. (1998). Spatial autocorrelation: a primer. Journal of housing economics, 7 (4), 304–327.

Efron, B. and Hastie, T. (2016). Computer age statistical inference. Vol. 5. : Cambridge University Press.

Ertz, O., Rey, S. J. and Joost, S. (2014). The open source dynamics in geospatial research and education. Journal of Spatial Information Science, 2014 (8), 67–71.

Fotheringham, A. S. (1997). Trends in quantitative methods i: stressing the local. Progress in Human Geography, 21 (1), 88–96.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. The incorporated statistician, 5 (3), 115–146.

Getis, A. (2008). A history of the concept of spatial autocorrelation: A geographer's perspective. Geographical Analysis, 40 (3), 297–309.

Getis, A. (2010). Spatial autocorrelation. In: Handbook of applied spatial analysis, pp. 255–278.

Getis, A. and Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. Geographical analysis, 36 (2), 90–104.

Goodchild, M. F. (1986). Spatial autocorrelation. Vol. 47. : Geo Books.

Grubesic, T. H., Wei, R. and Murray, A. T. (2014). Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. Annals of the Association of American Geographers, 104 (6), 1134–1156.

Han, J., Kamber, M. and Tung, A. K. (2001). Spatial clustering methods in data mining. Geographic data mining and knowledge discovery,, 188–217.

Han, J., Lee, J.-G. and Kamber, M. (2009). An overview of clustering methods in geographic data analysis. Geographic data mining and knowledge discovery,, 149–187.

Hesse, B. W., Moser, R. P. and 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.

Kaufman, L. and Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. Vol. 344. : John Wiley & Sons.

Keim, D. A., Mansmann, F., Schneidewind, J. and Ziegler, H. (2006). Challenges in visual data analysis. In: Tenth International Conference on Information Visualisation (IV'06). IEEE, pp. 9–16.

Kettenring, J. R. (2006). The practice of cluster analysis. Journal of classification, 23 (1), 3–30.

Lee, Y.-J. (2014). Social vulnerability indicators as a sustainable planning tool. Environmental Impact Assessment Review, 44, 31–42.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? Ecology, 74 (6), 1659–1673.

Legendre, P. and Fortin, M. J. (1989). Spatial pattern and ecological analysis. Vegetatio, 80 (2), 107–138.

Lloyd, C. D. (2010). Local models for spatial analysis. : CRC press.

Lu, C.-T., Chen, D. and Kou, Y. (2003). Algorithms for spatial outlier detection. In: Third IEEE International Conference on Data Mining. IEEE, pp. 597–600.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika, 37 (1/2), 17–23.

Ord, J. K. (2004). Spatial processes. Encyclopedia of Statistical Sciences,.

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

Oxoli, D. (2019). Exploratory approaches in spatial association analysis: methods, complements, and open gis tools development. Ph.D. thesis, Politecnico di Milano, Italy.

Oxoli, D., Prestifilippo, G., Bertocchi, D. and Zurbarán, M. (2017). Enabling spatial autocorrelation mapping in qgis: The hotspot analysis plugin. Geoingegneria Ambientale e Mineraria, 151 (2), 45–50.

Phipson, B. and 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).

Rey, S. J. (2001). Spatial empirics for economic growth and convergence. Geographical analysis, 33 (3), 195–214.

Schabenberger, O. and Gotway, C. A. (2017). Statistical methods for spatial data analysis. : Chapman and Hall/CRC.

Scrucca, L. and et al. (2005). Clustering multivariate spatial data based on local measures of spatial autocorrelation. Quaderni del Dipartimento di Economia, Finanza e Statistica, 20 (1), 11.

Smit, B. and Wandel, J. (2006). Adaptation, adaptive capacity and vulnerability. Global environmental change, 16 (3), 282–292.

Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., Negrini, R., Landguth, E., Jones, M. R., Consortium, N. and et al. (2017). High performance computation of landscape genomic models including local indicators of spatial association. Molecular ecology resources, 17 (5), 1072–1089.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. Economic geography, 46 (sup1), 234–240.

Tse, R. Y. (2002). Estimating neighbourhood effects in house prices: towards a new hedonic model approach. Urban studies, 39 (7), 1165–1180.

Tukey, J. (1977). Exploratory data analysis princeton, ed.

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

Wiegand, T. and Moloney, K. A. (2013). Handbook of spatial point-pattern analysis in ecology. : Chapman and Hall/CRC.

Wulder, M. and Boots, B. (1998). Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the getis statistic. International Journal of Remote Sensing, 19 (11), 2223–2231.

Xu, R. and Wunsch, D. (2008). Clustering. Vol. 10. : John Wiley & Sons.

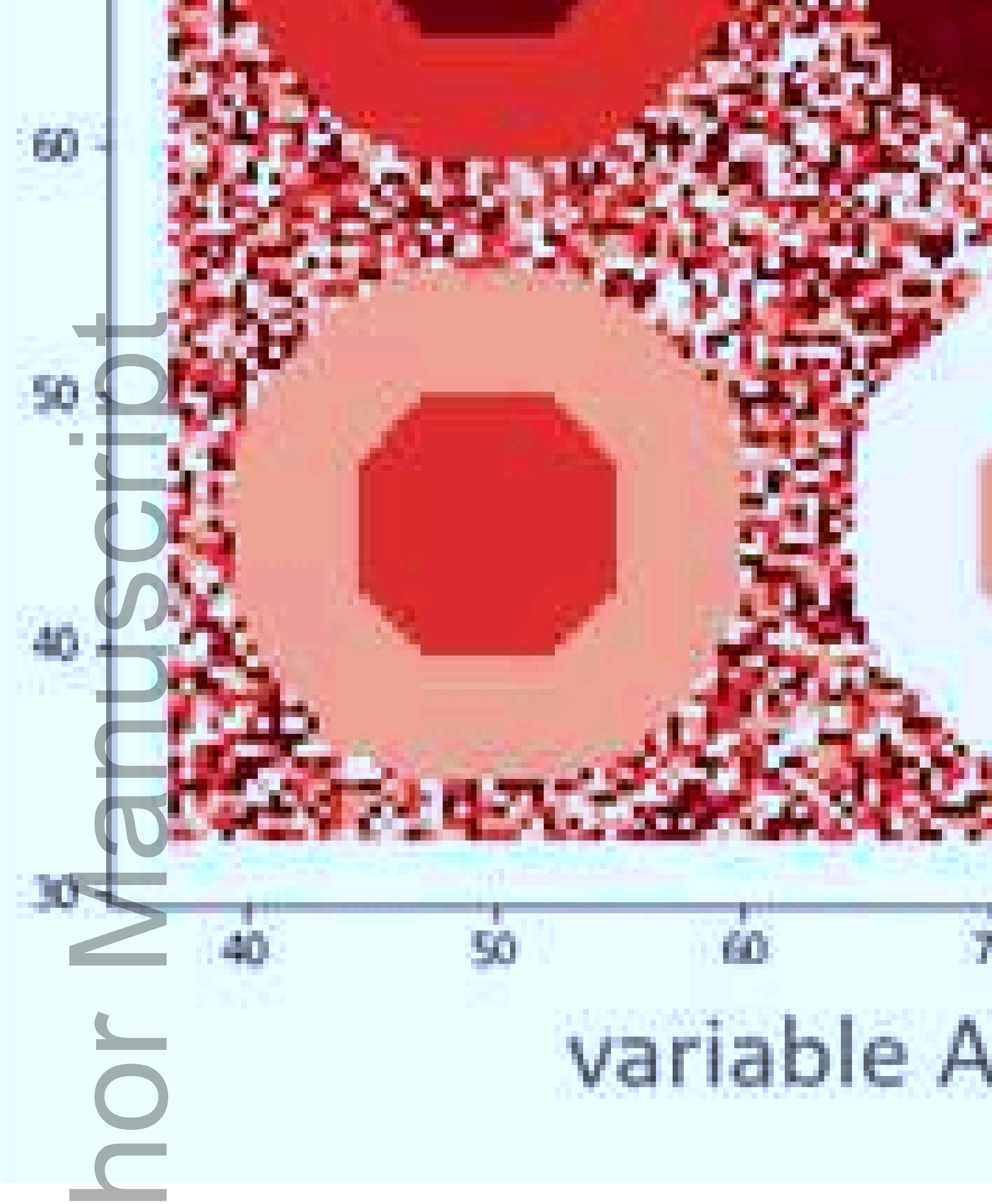


FIGURE 2 Synthetic multivariate spatial dataset. Sampling distributions of the simulated variables (a), schematic of the arbitrarily defined sub-regions (1,2,3,4,5,6,7 and 8) with the associated quartile intervals (b), and choropleth maps (c) of the four simulated variables (standardized values). The visualization style is based on the quartile breaks of the each variable distribution.

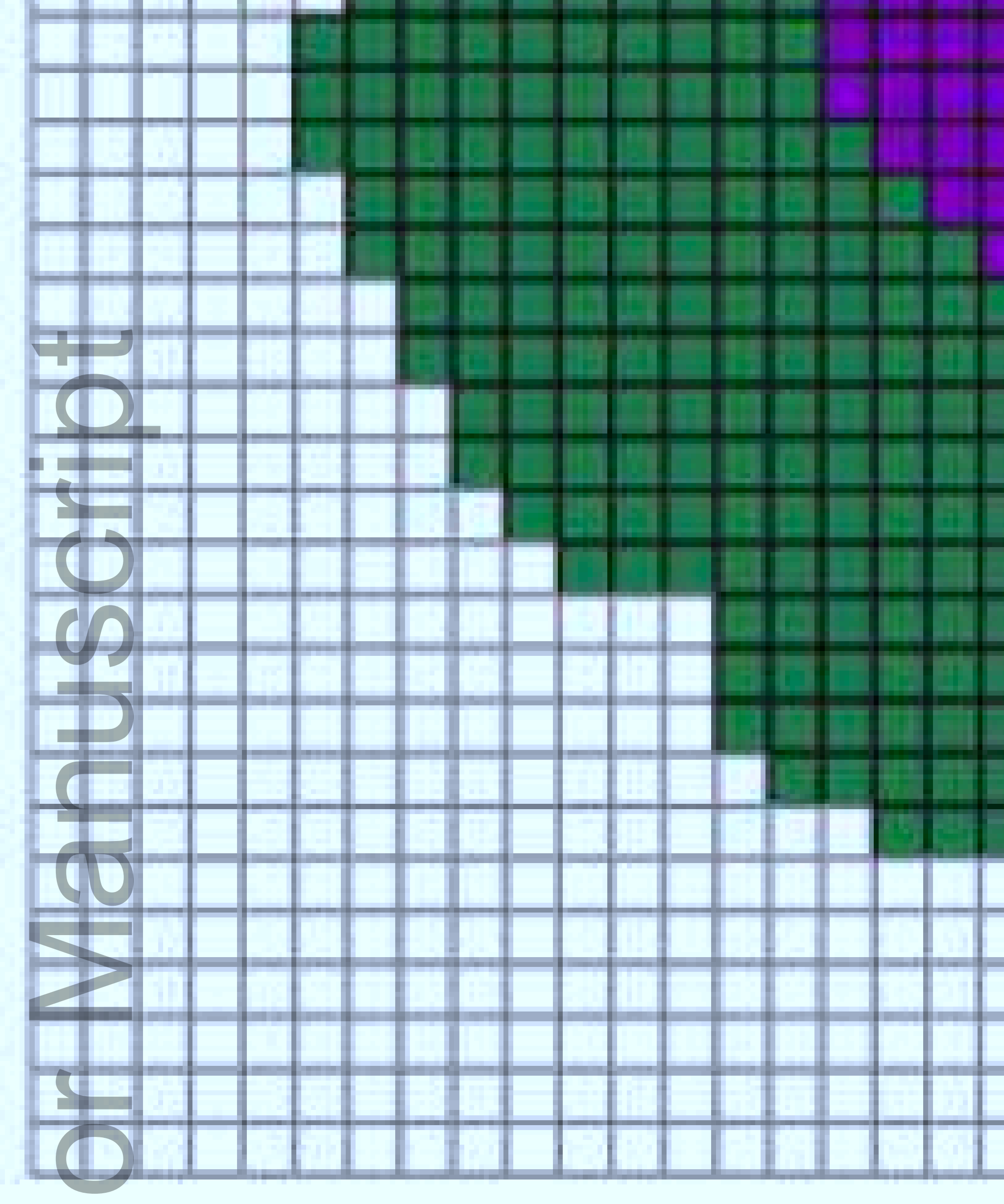


FIGURE 3 Multivariate clusters and outliers map from the computation of the multivariate local Geary's c and the Mm classification on the synthetic dataset.

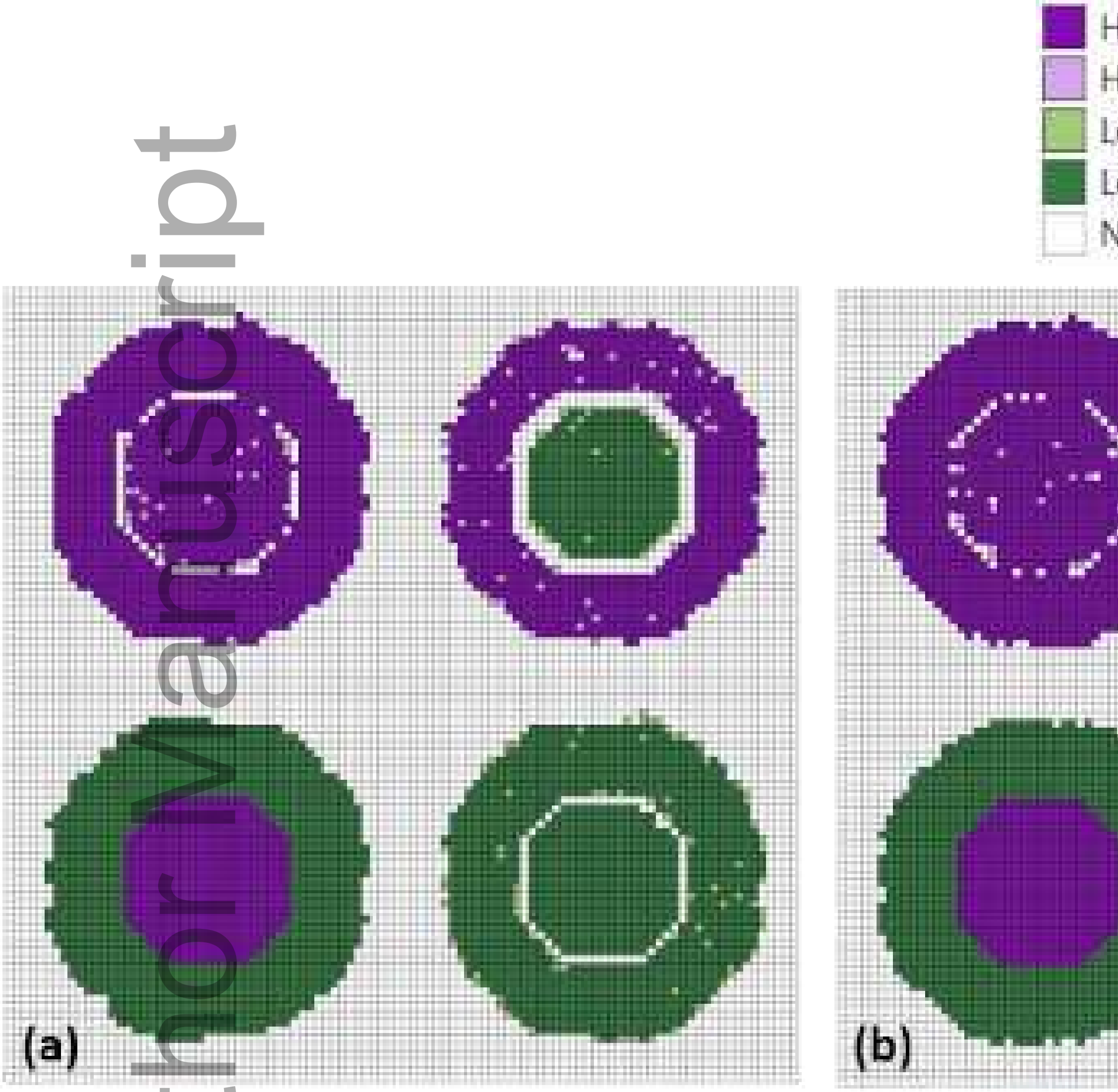


FIGURE 4 Multivariate clusters and outliers maps from the computation of the multivariate local Geary's c and the Mm classification on the synthetic dataset using different spatial weights matrices: Queen contiguity (a), rook contiguity (b) and Kernel Gaussian weights (c).

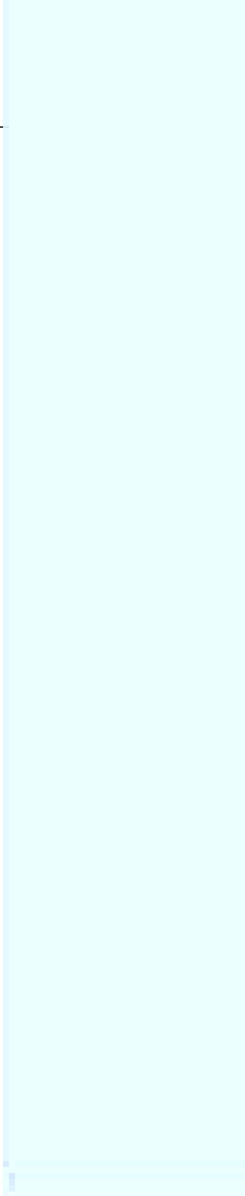


FIGURE 5 Choropleth map of the three considered indexes (a). The visualization style is based on the quartile breaks of each index distribution (standardized values). Connectivity graph based on the queen contiguity rule (b). The centroid of each census tract (black dot) is used to represent the corresponding polygon. Each line corresponds to a neighbour relation. Histogram of the neighbours distribution (c).

FIGURE 6 Resulting maps from the computation of the univariate local Moran's I for the three selected social vulnerability indexes. Maps are computed by means of the GeoDa software. The default color schema provided by the software is maintained for the clusters and outliers visualization.

FIGURE 7 Resulting map from the computation of the multivariate local Geary's c and the Mm classification for the three selected social vulnerability indexes.

References

Anderberg, M. R. (2014). *Cluster analysis for applications: Probability and mathematical statistics.* Cambridge, MA: Academic Press.

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis, 27*(2), 93-115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In M. M. Fischer (Ed.), *Spatial analytical perspectives in GIS* (pp. 111-125). London, UK: Taylor & Francis.

Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. In P. A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind (Eds.), *Geographical Information Systems: principles, techniques, management and applications* (Vol. 1, pp. 251-264). New York, NY: John Wiley & Sons.

Anselin, L. (2019). A local indicator of multivariate spatial association: Extending Geary's c. *Geographical Analysis, 51*(2), 133-150.

Anselin, L., & Getis, A. (1992). Spatial statistical analysis and geographic information systems. *Annals of Regional Science, 26*(1), 19-33.

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

Barrell, J., & Grant, J. (2013). Detecting hot and cold spots in a seagrass landscape using local indicators of spatial association. *Landscape Ecology, 28*(10), 2005-2018.

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), 57*(1), 289-300.

Besag, J., & Diggle, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 26*(3), 327-333.

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

Boots, B. (2003). Developing local measures of spatial association for categorical data. *Journal of Geographical Systems, 5*(2), 139-160.

Brovelli, M. A., Minghini, M., Moreno-Sanchez, R., & Oliveira, R. (2017). Free and open source software for geospatial applications (FOSS4g) to support future Earth. *International Journal of Digital Earth, 10*(4), 386-404.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial non-stationarity. *Geographical analysis, 28*(4), 281-298.

Cliff, A., & Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis, 4*(3), 267-284.

Cressie, N. (1990). The origins of kriging. *Mathematical Geology, 22*(3), 239-252.

Cromley, R. G., Hanink, D. M., & Bentley, G. C. (2014). Geographically weighted colocation quotients: specification and application. *Professional Geographer, 66*(1), 138-148.

Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology & Biogeography, 16*(2), 129-138.

Dray, S., & Jombart, T. (2011). Revisiting Guerry's data: Introducing spatial constraints in multivariate analysis. *Annals of Applied Statistics, 5*(4), 2278-2299.

Dubin, R. A. (1998). Spatial autocorrelation: A primer. *Journal of Housing Economics, 7*(4), 304-327.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference.* Cambridge, UK: Cambridge University Press.

Ertz, O., Rey, S. J., & Joost, S. (2014). The open source dynamics in geospatial research and education. *Journal of Spatial Information Science, 8,* 67-71.

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

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician, 5*(3), 115-146.

Getis, A. (2008). A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis, 40*(3), 297-309.

Getis, A. (2010). Spatial autocorrelation. In M. M. Fischer & A. Getis (Eds.), *Handbook of applied spatial analysis: Software tools, methods and applications* (pp. 255-278). Berlin, Germany: Springer.

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

Goodchild, M. F. (1986). *Spatial autocorrelation.* Norwich, UK: Geo Books.

Grubesic, T. H., Wei, R., & Murray, A. T. (2014). Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers, 104*(6), 1134-1156.

Han, J., Kamber, M., & Tung, A. K. (2001). Spatial clustering methods in data mining. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (pp. 188-217). Boca Raton, FL: Chapman & Hall/CRC.

Han, J., Lee, J.-G. and Kamber, M. (2009). An overview of clustering methods in geographic data analysis. In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery* (pp. 149-187). Boca Raton, FL: Chapman & Hall/CRC.

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

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis.* New York, NY: John Wiley & Sons.

Keim, D. A., Mansmann, F., Schneidewind, J., & Ziegler, H. (2006). Challenges in visual data analysis. In *Proceedings of the 10th International Conference on Information Visualisation, London, UK* (pp. 9-16). Piscataway, NJ: IEEE.

Kettenring, J. R. (2006). The practice of cluster analysis. *Journal of Classification, 23*(1), 3-30.

Lee, Y.-J. (2014). Social vulnerability indicators as a sustainable planning tool. *Environmental Impact Assessment Review, 44,* 31-42.

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

Legendre, P., & Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Vegetatio, 80*(2), 107-138.

Lloyd, C. D. (2010). *Local models for spatial analysis.* Boca Raton, FL: CRC Press.

Lu, C.-T., Chen, D. and Kou, Y. (2003). Algorithms for spatial outlier detection. In *Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, FL* (pp. 597-600). Piscataway, NJ: IEEE.

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

Ord, J. K. (2004). Spatial processes. In S. Kotz, C. B. Read, N. Balakrishnan, & B. Vidakovic (Eds.), *Encyclopedia of statistical sciences* (pp. 1-6). New York, NY: John Wiley & Sons.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis, 27*(4), 286-306.

Oxoli, D. (2019). *Exploratory approaches in spatial association analysis: Methods, complements, and open GIS tools development* (Unpublished Ph.D. dissertation). Politecnico di Milano, Italy.

Oxoli, D., Prestifilippo, G., Bertocchi, D., & Zurbarán, M. (2017). Enabling spatial autocorrelation mapping in QGIS: The hotspot analysis plugin. *Geoingegneria Ambientale e Mineraria, 151*(2), 45-50.

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 & Molecular Biology, 9,* 39.

Rey, S. J. (2001). Spatial empirics for economic growth and convergence. *Geographical Analysis, 33*(3), 195-214.

Schabenberger, O., & Gotway, C. A. (2017). *Statistical methods for spatial data analysis.* Boca Raton, FL: Chapman and Hall/CRC.

Scrucca, L. (2005). Clustering multivariate spatial data based on local measures of spatial autocorrelation. *Quaderni del Dipartimento di Economia, Finanza e Statistica, 20*(1), 11.

Smit, B., & Wandel, J. (2006). Adaptation, adaptive capacity and vulnerability. *Global Environmental Change, 16*(3), 282-292.

Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., … Joost, S. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources, 17*(5), 1072-1089.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*(Suppl. 1), 234-240.

Tse, R. Y. (2002). Estimating neighbourhood effects in house prices: Towards a new hedonic model approach. *Urban Studies, 39*(7), 1165-1180.

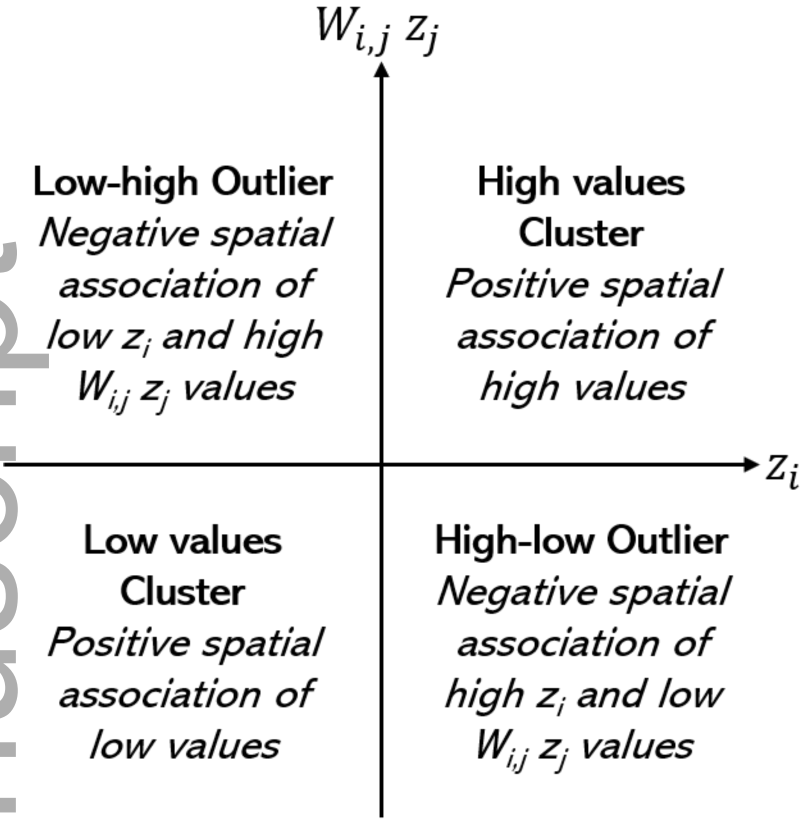
Tukey, J. W. (1977). *Exploratory data analysis.* Boston, MA: Addison-Wesley.

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

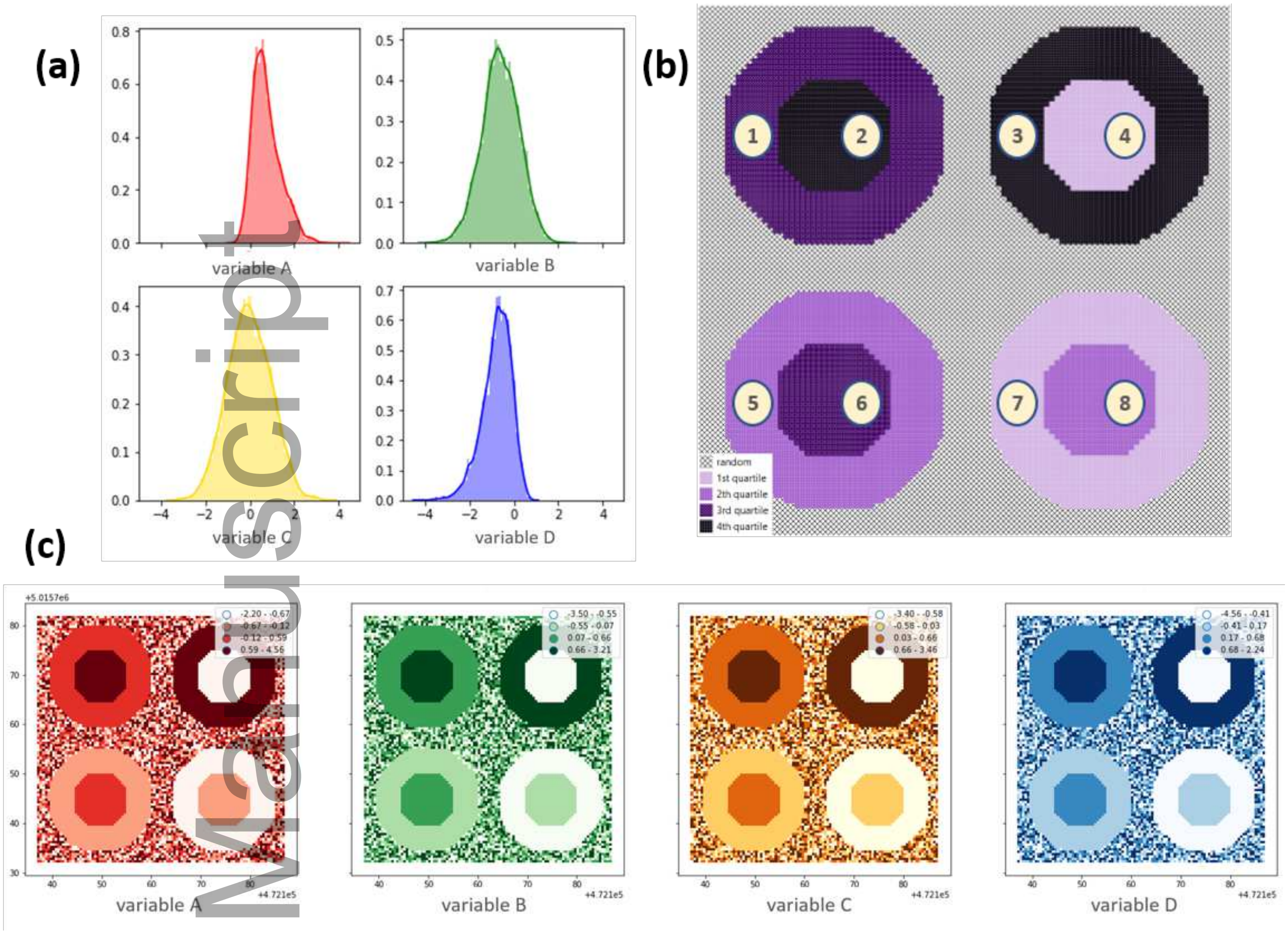
Wiegand, T., & Moloney, K. A. (2013). *Handbook of spatial point-pattern analysis in ecology.* Boca Raton, FL: Chapman & Hall/CRC.

Wulder, M., & Boots, B. (1998). Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the Getis statistic. *International Journal of Remote Sensing, 19*(11), 2223-2231.

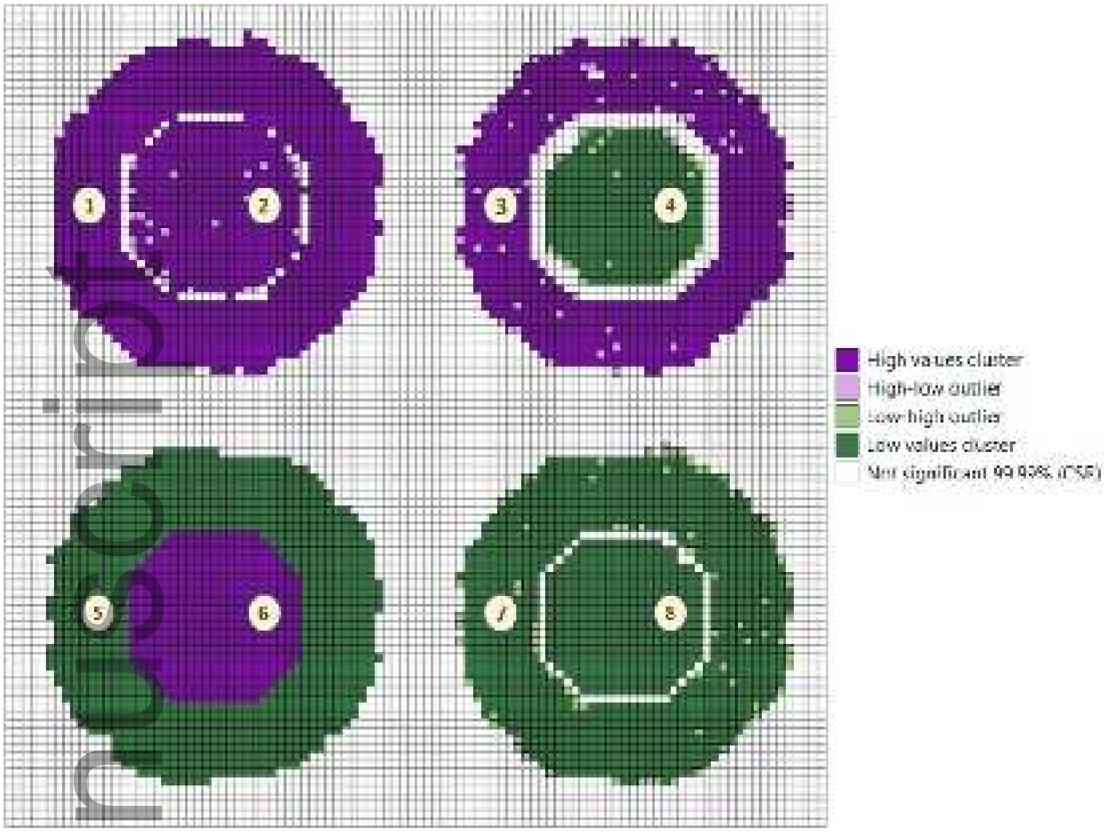
Xu, R., & Wunsch, D. (2008). *Clustering.* New York, NY: John Wiley & Sons.


$W_{i,j}\ z_j$
Low-high Outlier
Negative spatial association of low z_i and high $W_{i,j}\ z_j$ values
High values Cluster
Positive spatial association of high values
z_i
Low values Cluster
Positive spatial association of low values
High-low Outlier
Negative spatial association of high z_i and low $W_{i,j}\ z_j$ values

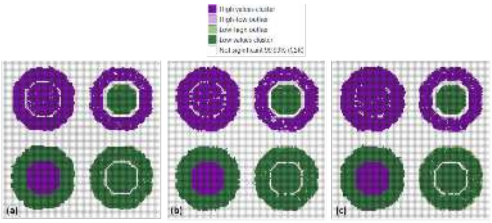
tgis_12639_f1.png



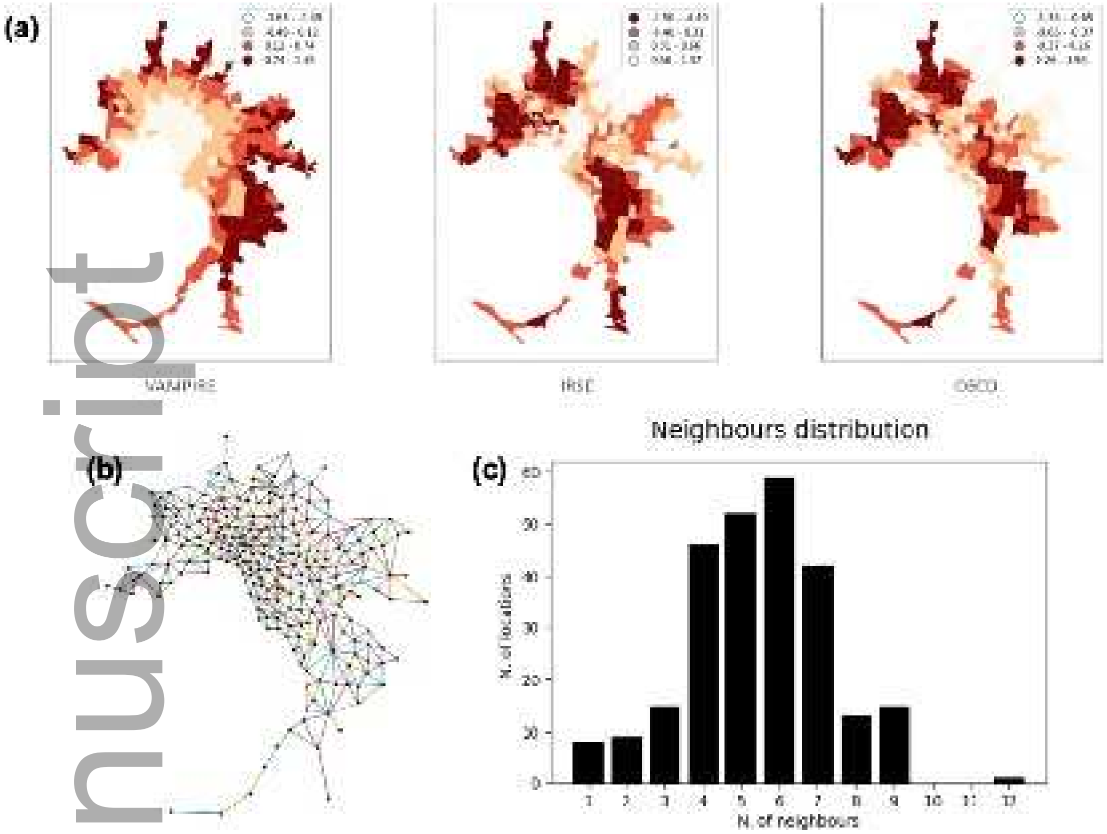
tgis_12639_f2.png



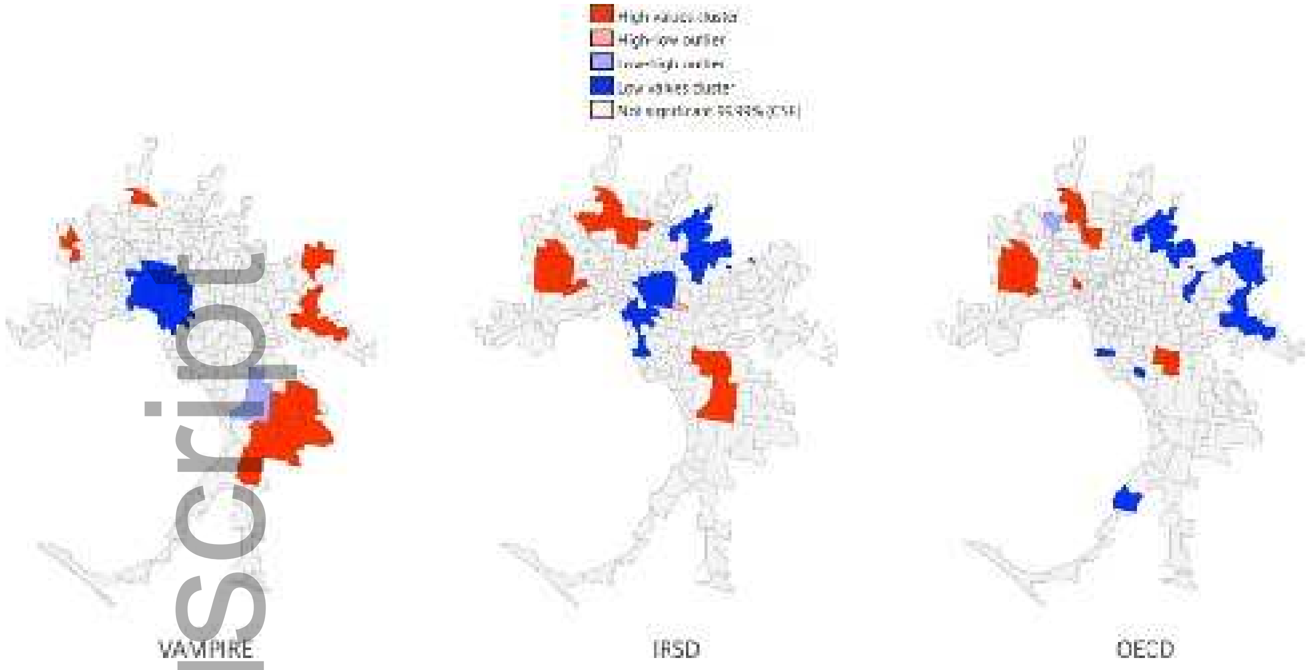
tgis_12639_f3.png



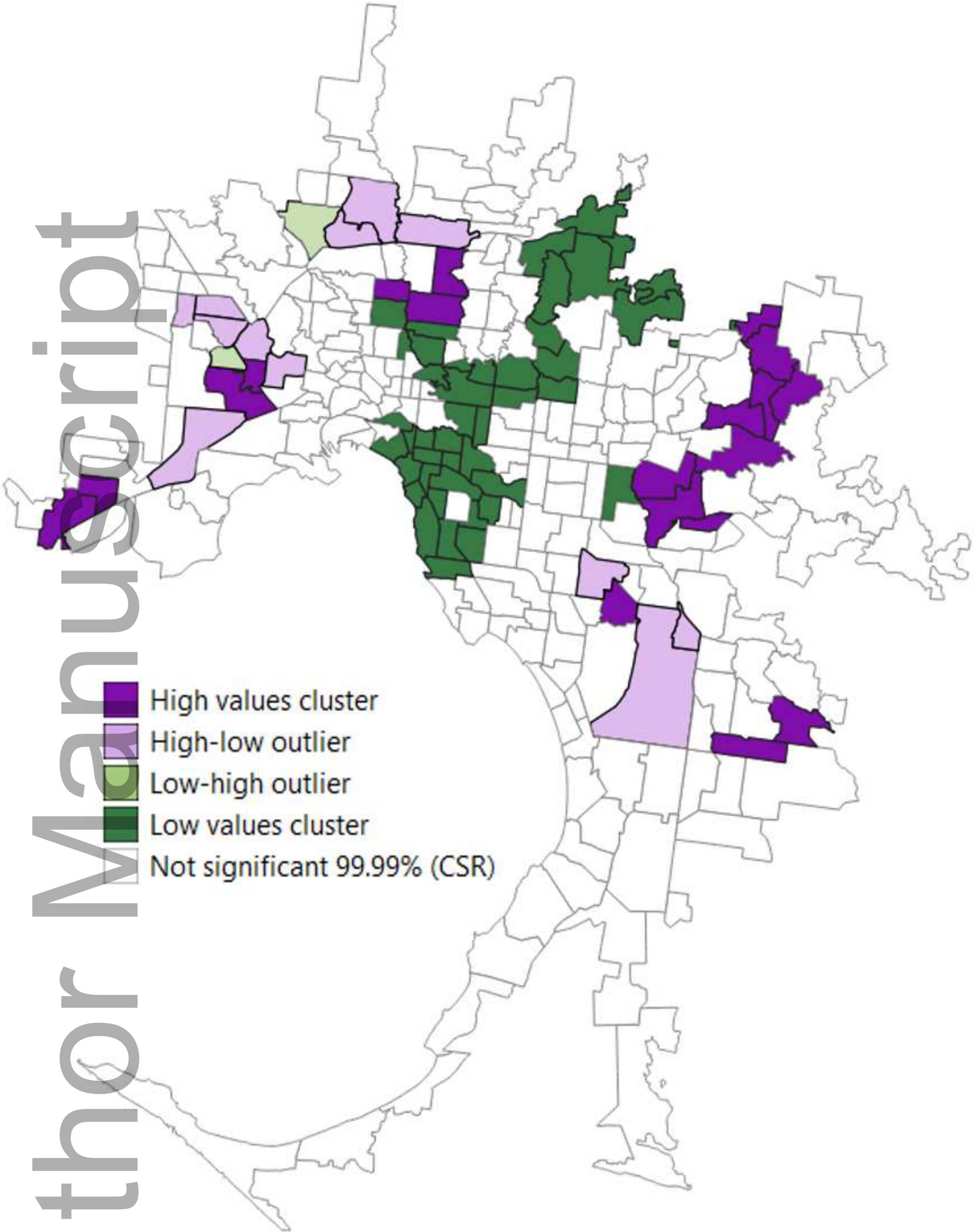
tgis_12639_f4.png



tgis_12639_f5.png



tgis_12639_f6.png



tgis_12639_f7.png

Author/s:
Oxoli, D;Sabri, S;Rajabifard, A;Brovelli, MA

Title:
A classification technique for local multivariate clusters and outliers of spatial association

Date:
2020-10-01

Citation:
Oxoli, D., Sabri, S., Rajabifard, A. & Brovelli, M. A. (2020). A classification technique for local multivariate clusters and outliers of spatial association. Transactions in Geographic Information Systems (GIS), 24 (5), pp.1227-1247. https://doi.org/10.1111/tgis.12639.

Persistent Link:
http://hdl.handle.net/11343/276882