Enabling spatial autocorrelation mapping in QGIS: The Hotspot Analysis Plugin

D. Oxoli*
G. Prestifilippo*
D. Bertocchi**
M. Zurbarán***

* Department of Civil and Environmental Engineering, Politecnico di Milano, Italy
** Department of Economics, University Ca' Foscari of Venice, Italy
*** Department of Systems Engineering, Universidad del Norte, Barranquilla, Colombia

The analysis of spatial autocorrelation is a fundamental tool for the understanding of all the physical as well as anthropological processes which naturally take place within the geographical space, and which cannot be studied independently from it.

The deployment of statistical techniques for investigating spatial autocorrelation has brought valuable results within manifold research fields ranging from the natural sciences to the socio-economic sciences. Moreover, the affinity between cartography and this kind of analysis has raised particular interest among GIS users as well as developers. This has led to the inclusion of many modules dedicated to the spatial autocorrelation mapping within both proprietary GIS software suites as well as free and open source programming libraries.

Nevertheless, specific functionalities for spatial autocorrelation mapping have not yet been formally included – through dedicated user interfaces – within the most popular free and open source GIS software, such as QGIS.

We present here the Hotspot Analysis Plugin, an experimental QGIS plugin – dedicated to the spatial autocorrelation mapping – based on the free and open source Python library PySAL (Python Spatial Analysis Library). Together with the technical specifications, two relevant examples of the plugin usage – connected to real case studies – are reported. These are: the detection of significant variations in soil consumption for the Lombardy Region (northern Italy) and the spatial correlation analysis of performance indicators characterizing Airbnb™ lodgings for the city of Venice (Italy).

***Keywords**: Hotspot Analysis, LISA, QGIS, Python, FOSS4G.*

Hotspot Analysis: sviluppo di applicativi per la mappatura dell'autocorrelazione spaziale in QGIS. *L'analisi dell'autocorrelazione spaziale è uno strumento indispensabile per la comprensione di tutti quei processi, naturali o antropici, che si manifestano in un dato territorio e che non possono essere studiati distintamente da esso.*

L'utilizzo di tecniche statistiche per l'analisi dell'autocorrelazione spaziale ha condotto a risultati importanti in svariati ambiti di ricerca che spaziano delle scienze naturali alle scienze sociali ed economiche. Inoltre, l'affinità tra questo tipo di analisi e la cartografia ha attirato una crescente attenzione da parte della comunità di utilizzatori e sviluppatori di software GIS. Questo ha portato alla nascita di moduli dedicati alla mappatura dell'autocorrelazione spaziale all'interno dei più moderni software GIS proprietari nonché in diverse librerie di programmazione open source.

Tuttavia, specifiche funzionalità per la mappatura dell'autocorrelazione spaziale non sono ad oggi ufficialmente integrate – tramite interfacce dedicate – nei più famosi software GIS open source, quale ad esempio QGIS.

Nel presente studio viene presentato un plugin sperimentale di QGIS – dedicato alla mappatura dell'autocorrelazione spaziale – chiamato Hotspot Analysis, basato sulla libreria open source PySAL (Python Spatial Analysis Library). Insieme alle caratteristiche tecniche, vengono riportati due esempi rilevanti di utilizzo del plugin – relativi a casi di studio reali – quali l'identificazione di variazioni significative del consumo di suolo per la regione Lombardia (IT) e la correlazione spaziale tra indicatori di performance, caratterizzanti le strutture ricettive di Airbnb™ per la città di Venezia (IT).

***Parole chiave**: Hotspot Analysis, LISA, QGIS, Python, FOSS4G.*

1. Introduction

Spatial autocorrelation can be defined as the degree to which characteristics at a certain location are similar to – or different from – those nearby. This concept is central to both Geography and GIS (Geographic Information System) science as it represents the confirmation of the Tobler's first law of Geography: "*Everything is related to everything else, but near things are more related than distant things*." (Tobler, 1970).

Spatial autocorrelation is the main subject of the Exploratory Spatial Data Analysis (ESDA) which identifies a collection of statistical and graphical techniques to describe and visualize spatial distributions, highlight atypical locations or outliers, discover patterns and suggest different spatial regimes and other forms of spatial instability (Anselin, 1999). In the past decades, ESDA brought valuable answers to different research fields such as epidemiology (e.g. Jones *et al.*, 2008), criminology (e.g. Andresen, 2006), disaster risk management (e.g. Zhou *et al.*, 2014), wildlife biology (e.g. Bone *et al.*, 2013), etc. by leading to an increasing consideration of these techniques among both GIS scientists and GIS software developers.

This success is mainly due to the adaptability of ESDA tools to a broad set of environmental data as well as socio-economic data, with the only constrain of providing that data refer to a specific location and their spatial relationship really matters for accomplishing the analysis objec-

For understanding the figures in a better way, please refer to the online version of the article, available on the GEAM journal website.

Per una migliore comprensione delle figure, si rimanda alla versione online disponibile sul sito internet della rivista GEAM.

tive. In general terms, this latter represents the rule in dealing with geospatial data.

Of particular interest for this work is a set of ESDA statistical tools called Local Indicators of Spatial Association (LISA) (Anselin, 1995). These statistics aim to identify the presence of local autocorrelation patterns (e.g. hotspots) in the spatial arrangement of a given variable. Due to the explicit consideration of spatial relationship between the data, results from LISA statistics lend themselves perfectly to be disclosed through maps, which is the reason that explains the popularity of LISA within the GIS context.

In fact, LISA relies on various software implementations which is a strong triggering factor for spreading the usage of these techniques. LISA mapping functionalities are included in the proprietary software ArcGIS (https://www.arcgis.com), into different stand-alone Free and Open Source Software (FOSS) dedicated to geospatial data analysis such as GeoDa (http://geodacenter.github.io) as well as into programming libraries like the R-spdep (https://cran.r-project.org/web/packages/spdep) and the Python-PySAL (Rey and Anselin, 2009).

Within the most popular FOSS GIS, like QGIS (www.qgis.org), the access to LISA mapping capabilities is currently enabled only through command line while dedicated plugins have not been formally made available yet.

According to this latter consideration, we present here the Hotspot Analysis Plugin, an experimental QGIS Python plugin aimed both to facilitate the access to LISA mapping tools for users with no advanced programming skills – exploiting the user friendly QGIS environment – as well as to contribute to the growth of the mapping capabilities of this FOSS GIS software.

Focused on the key points mentioned above, the rest of the paper is structured as follows. In the next section the technical characteristics of the plugin are described. This is followed by the two examples of plugin usage within real case studies. The paper closes outlining the main conclusions and perspective for the plugin implementation work.

2. The Hotspot Analysis Plugin

QGIS is much more than a GIS software; it is also a geospatial programming environment, which enables users to build geospatial applications using Python. Moreover – being QGIS a FOSS – it relies on a widespread community of developers which collaborate to improve system capabilities, by adding processing functionalities through plugins, as well as to maintain and upgrade the system core.

While QGIS core itself is written in C++, it includes extensive support for Python programming (Westra, 2014). QGIS Python plugins share a common structure which is based on the Python bindings of the Qt framework (i.e. PyQt). Basic operations such as load data sources into layers, manipulate and export maps, etc. are available directly from PyQGIS which is the Python package included in the QGIS default installation.

2.1. Plugin Implementation

The Hotspot Analysis Plugin is mainly based on the Exploratory Spatial Data Analysis (ESDA) module of PySAL and PyQGIS, providing a simplified interface to run LISA tools starting from vector layers. The PySAL library has not yet made available within the QGIS default installation. Therefore, the prior installation into QGIS of PySAL and its dependencies (i.e. the libraries NumPy and SciPy) is required. Installation guidelines are provided, together with the plugin source code, on GitHub (https://github.com/danioxoli/HotSpotAnalysis_Plugin). The stable release of the plugin is available on the official QGIS Python Plugins Repository (https://plugins.qgis.org/plugins/HotspotAnalysis)

2.2. Plugin Functionalities

The Hotspot Analysis Plugin requires as input a projected shapefile of points or polygon with at least one numerical attribute associated to each geometry (Fig. 1b). This latter depicts the variable for which LISA is computed. Examples of numerical attribute are: census data, average house prices, earthquake intensity, etc. 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 an adjacency – or spatial weights – binary matrix, exploiting the dedicated PySAL functionalities. The spatial arrangement of the geometries is retrieved directly from the input shapefile. For what it concerns pointwise geometries, the matrix is created using a fixed distance band (expressed with the same unit of measure of the projected coordinate system of the input shapefile). Alternatively, the matrix can be created using the K-Nearest Neighbours (KNN) approach, which enables to define a relation for any point of the dataset with its K nearest points (where K value is set by the user). For shapefiles of polygons, a first order queen's case contiguity matrix is used (i.e. edge and/or corners contiguity).

LISA computations implemented into the Hotspot Analysis Plugin are: a) the Getis-Ord Gi* statistic (Ord and Getis, 1995), b) the Local Moran's I statistic (Anselin, 1995), and c) the Bivariate Local Moran statistic (Wartenberg, 1985). The plugin user interface (Fig. 1a) allows to select one among the three aforementioned LISA for each plugin run. The selected LISA is computed starting from the spatial weights matrix and the specified numerical attribute at any location of the dataset.

The general formulation for the a) Getis-Ord Gi* is given by the equations (1):

$$Gi^* = \frac{\Sigma_j w_{ij} x_j}{\Sigma_j x_j}, i \in j \qquad (1)$$

where x_j is the value of the attribute at the j^{th} location and w_{ij} is the element of the spatial weights matrix correspondent to the i^{th} and j^{th} relationship.

The b) Local Moran's I is defined in equation (2):

$$I_i = z_i \Sigma_j w_{ij} z_i,\ i \neq i \qquad (2)$$

where z_i and z_j are the standardized values of the attribute at the i^{th} and j^{th} locations and w_{ij} is the element of the spatial weights matrix correspondent to the i^{th} and j^{th} relationship.

The c) Bivariate Local Moran (i.e. the bivariate counterpart of the Local Moran's I) requires two numerical attributes to be assigned to each geometry. The general formulation is given by equation (3):

$$I^i_{kl} = z^i_k \Sigma_j w_{ij} z^j_l, i \neq j \qquad (3)$$

where z_i^k and z_j^l are the standardized values of the two attributes k and l at the i^{th} and j^{th} locations while w_{ij} is the element of the spatial weights matrix correspondent to the i^{th} and j^{th} relationship.

The plugin processing output consists of a copy of the input shapefile to which two new columns of the attribute table are added. These contain the normal standard variates (Z-scores) of the computed LISA and their statistical significance (p-values) related to the null hypothesis (i.e. complete spatial randomness) under normality assumption.

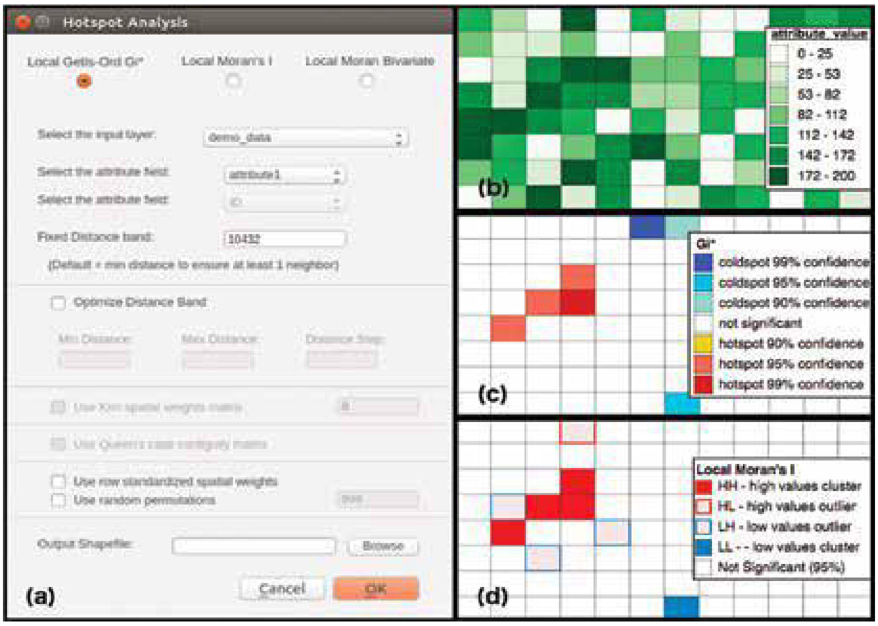
The output layer is displayed with an automatic style that combines Z-scores and p-values allowing an intuitive visualization of the detected local spatial clusters (Fig. 1c-1d).

For what it concerns Getis-Ord Gi*, a positive and statistically significant Z-score indicates a cluster of high values (hotspot). A negative and statistically significant Z-score indicates a cluster of low values (coldspot). With respect to 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 (Anselin, 1999) which are included into a separate column of the attribute table. The q-values depict presence of clusters or outliers within the dataset. Interpretations of clusters/ outliers differs between the Local Moran's I and the Local Moran Bivariate. For this latter, clusters and outliers are detected for the base attribute k with respect to the surrounding values of the second attribute l. This is exactly the meaning of bivariate spatial autocorrelation of which the Local Moran Bivariate is a measure.

Plugin additional functionalities are: the selection of row standardized spatial weights matrix (instead of the default binary version), the computation of statistical significance using permutation approach (instead of normality assumption), and a semi-automatic optimization for the selection of the fixed distance band. This latter allows user to specified a range of possible distances to be test by the plugin. Selected distance is the one maximizing the Z-score of the Global Moran's I index (Anselin, 1995) for the dataset. A default distance band is suggested automatically for any specified input shapefile of points. This distance is the minimum distance that guarantees at least one neighbour to each point (Fig. 1a).

3. *Plugin application examples*

In this section are reported two meaningful examples on the usage of the Hotspot Analysis Plugin in the context of real ongoing rese-



Hotspot Analysis

● Local Getis-Ord Gi* ○ Local Moran's I ○ Local Moran Bivariate

Select the input layer: demo_data

Select the attribute field: attribute1

Select the attribute field: ID

Fixed Distance band: 10432

(Default = min distance to ensure at least 1 neighbor)

□ Optimize Distance Band

Min Distance: Max Distance: Distance Step:

□ Use Knn spatial weights matrix 8

□ Use Queen's case contiguity matrix

□ Use row standardized spatial weights

□ Use random permutations 999

Output Shapefile: Browse

Cancel OK

Fig. 1. Principali caratteristiche del plugin: (a) interfaccia grafica, (b) un possibile esempio di dato in ingresso basato su una griglia regolare, (c) stile automatico assegnato allo strato risultante dal calcolo di Getis-Ord Gi* e (d) dal calcolo di Local Moran's I.
Main plugin features: (a) user interface, (b) an example of input data based on a regular grid, (c) default output layer styling for the Getis-Ord Gi, and (d) for the Local Moran's I.*

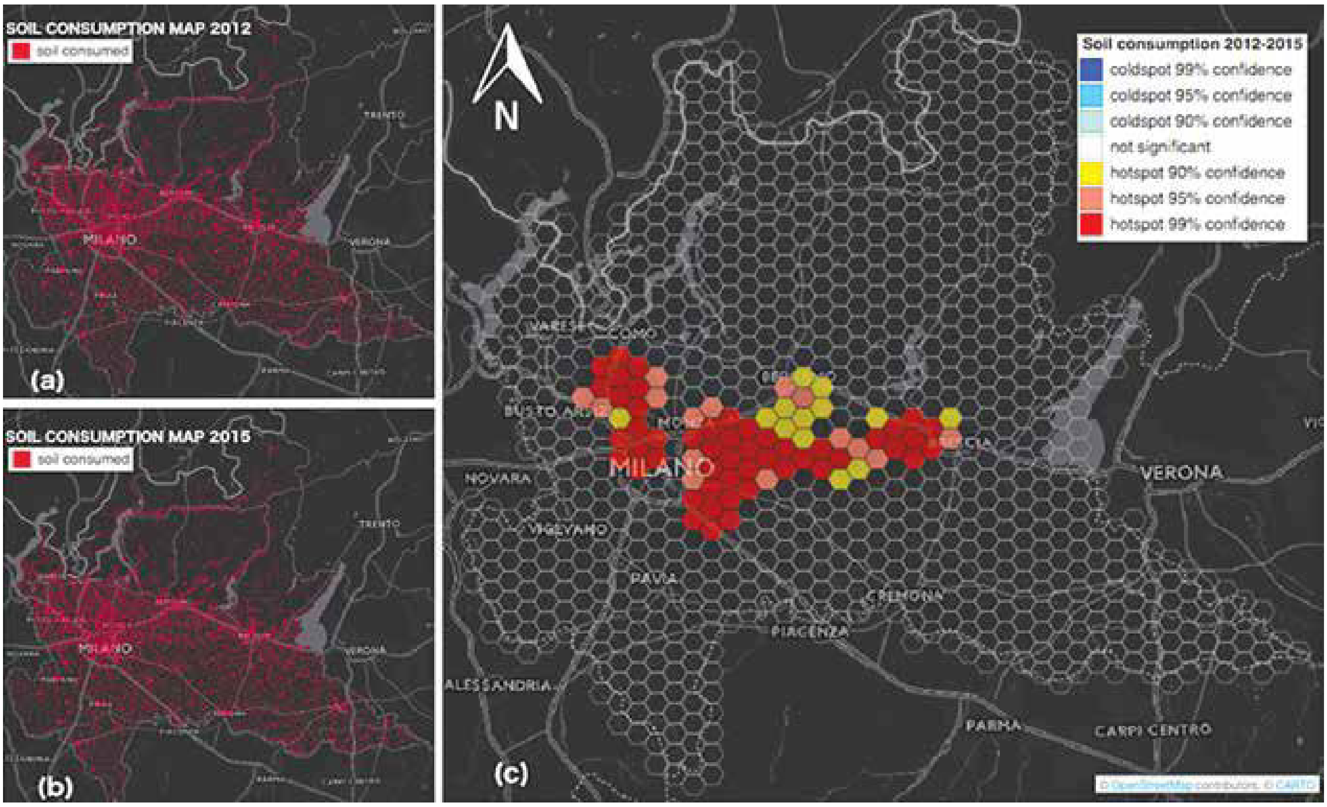


Fig. 2. Mappe di consumo del suolo riferite agli anni (a) 2012 e (b) 2015 per la regione Lombardia; lo strato risultante dal calcolo di Getis-Ord Gi* (c) mostra le aree in cui il consumo di suolo si è verificato in maniera significativa tra i due anni di riferimento.
Soil consumption maps related to the year (a) 2012 and (b) 2015 for the Lombardy Region (Northern Italy); the output layer resulting from the computation of the Getis-Ord Gi (c) shows the areas where soil consumption process significantly took place between the two reference years.*

arch work involving the authors. Examples regard the analysis of soil consumption maps (see Section 3.1) and the spatial autocorrelation of some Airbnb™ lodgings indicators (see Section 3.2). These have been chosen to demonstrate the effectiveness of adopting hotspot and cluster mapping – or more in general ESDA – for the analysis of heterogeneous geospatial data.

3.1. *Soil Consumption*

The term "soil consumption" describes the loss and the degradation of soil and lands which often imply loss of ecosystem services and – in turn – an increase of vulnerability to climate change (Keesstra *et al.*, 2016). Therefore, the understanding of this phenomenon plays a key role in the design of sustainable land management policies.

In Italy, soil consumption monitoring is performed by the Italian National Institute for Environmental Protection and Research (ISPRA), which provides an updated view of the soil consumption process through the periodic publication of soil consumption maps. Currently, two binary raster maps at a national scale, with a resolution of 10 m, are available (Fig. 2a-2b). These are related to the reference years 2012 and 2015.

In order to detect and locate significant losses of soil occurred between the two reference years, a raster map of the differences was computed by subtracting 2012 map from 2015 map. Thus, pixels containing positive differences (i.e. new areas affected by soil consumption) were aggregated over a regular grid which served as input layer to run the Hotspot Analysis Plugin.

In this case, the analysis variable was the number of these pixels per grid cell and the selected LISA was the Getis-Ord Gi*. This analysis was performed for the Lombardy Region (Northern Italy) by using a regular hexagonal grid. Results are included in Figure 2c. It can be observed that soil consumption hotspots are concentrated along the central part of the east-west axis, where the major cities are also located. Definitely, the resulting map is an effective tool to support the investigation of critical areas affected by the soil consumption process.

3.2. *Airbnb™ lodging analysis*

Airbnb™ is an online 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 col-

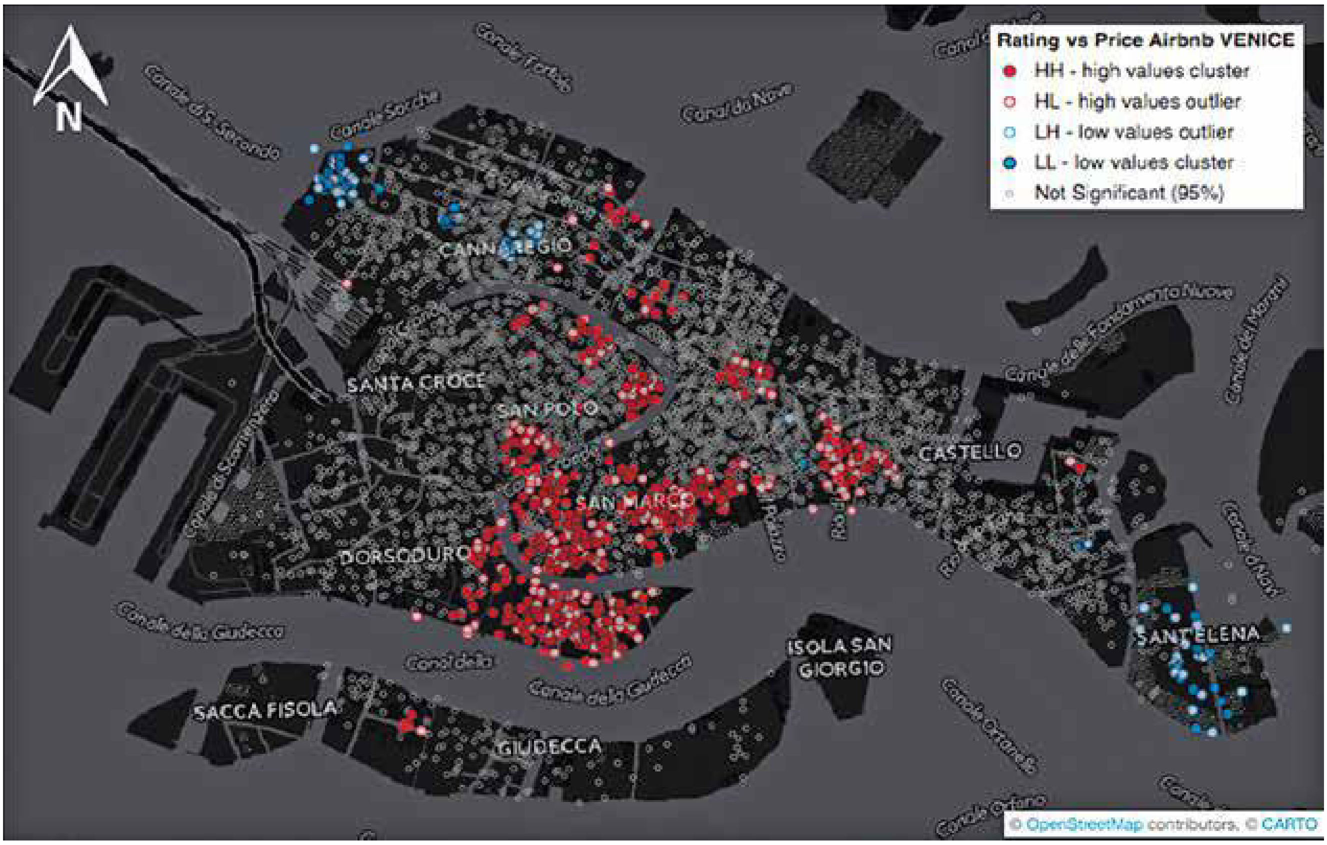


Fig. 3. Mappa risultante dal calcolo dell'indicatore di Moran locale bivariato per le strutture ricettive Airbnb™ di Venezia; clusters e valori anomali nella distribuzione spaziale delle valutazioni delle strutture sono evidenziati in relazione al loro prezzo medio per notte.

Resulting layer from the computation of the Local Moran Bivariate for the Airbnb™ lodgings of Venice; clusters and outliers in the spatial distribution of lodging ratings, with respect to their average price per night, are highlighted through a specific layer symbology.

lection/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 quality of the service provided by the Airbnb™ marketplace itself. This information has also a not negligible value in the fields of tourism economics and territory management (Van der Zee *et al.*, 2016). In fact, by considering data such as lodging prices, ratings, reviews, etc. in relation to their geographical space, it is possible to perform analyses on the territory attractiveness which provide valuable inputs for the implementation of proper territorial requalification, conservation as well as promotion policies.

The presented example focuses on the analysis of the spatial correlation between Airbnb™ lodging average prices and ratings for the city of Venice (IT). The input dataset includes ~ 4000 pointwise geometries describing lodgings locations and attributes. The selected LISA was the Local Moran Bivariate, computed for the lodging ratings (as base attribute) and the prices. Due to the significant variation of points density within the area of analysis, the spatial weights matrix was computed by using the KNN strategy, with K value set to 30. This to prevent an uneven distribution of spatial weights by ensuring enough neighbours to each lodging. According to Ord and Getis (1995), this is a good practice in order to assume normality for the significance test. The resulting map is shown in Figure 3. It can be observed that high values clusters (i.e. locations where high ratings are surrounded by high prices) are mostly scattered in the neighbourhood of the principal public square (i.e. St Mark's Square) and main sights and points of interest of the destination (i.e. museums, churches, etc.). Conversely, low values clusters are segregated in the peripheral areas. With this analysis, it is possible to clearly asses the differences in attractiveness among locations of the same city as well as in the Airbnb™ accommodation market.

4. Conclusions and further work

Potential applications of the Hotspot analysis – or more in general of ESDA – are broad and helpful for manifold scientific fields. The possibility to perform this kind of spatial analysis within QGIS represents a valuable asset to boost the use of this FOSS GIS among a larger and heterogeneous users' community. The inclusion of powerful spatial analysis libraries – such as PySAL – within the QGIS core libraries represents a meaningful objective in order to strengthen the capabilities of this software.

The Hotspot Analysis Plugin aims exactly to bridge the gap of geospatial analysis tools of QGIS with respect to the other commercial software and – at the same time – to empower the QGIS users' community with powerful additional functionalities.

Therefore, further improvements will focus on: the inclusion of others PySAL-ESDA tools in the plugin, the optimization of both the source code and the graphical user interface and the operational testing of the plugin in the context of more extensive research work.

References

Andresen, M.A., 2006. *Crime measures and the spatial analysis of criminal activity*. British Journal of Criminology. 46(2). pp. 258-285.

Anselin, L., 1995. *Local Indicators of Spatial Association – LISA. Geographical Analysis*, 27(2), pp. 93-115.

Anselin, L., 1999. *Interactive techniques and exploratory spatial data analysis*, in: Geographical Information Systems: Principles, Techniques, Management and Applications, (Eds.) Longley, P., Goodchild, M., D. Maguire, D., Rhind, D., Geoinformation Int, Cambridge, pp. 253-266.

Bone, C., Wulder, M.A., White, J.C., Robertson, C., Nelson, T.A., 2013. *A GIS-based risk rating of forest insect outbreaks using aerial overview surveys and the local Moran's I statistic*. Applied Geography. 40. pp. 161-170.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. *Global trends in emerging infectious diseases*. Nature. 451(7181). pp. 990-993.

Keesstra, S.D., Quinton, J.N., van der Putten, W.H., Bardgett, R.D., Fresco, L.O., 2016. *The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals*. Soil. 2(2). pp. 111.

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

Rey, S., Anselin, L., 2009. *PySAL: A Python library of spatial analytical methods*. Spatial Analysis: Software Tools, Methods. 37(2007). pp. 175-193.

Tobler, W.R., 1970. *A computer movie simulation urban growth in Detroit region*. Economic Geography. 46. pp. 234-240.

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 Kraków*, in: Proceedings of TCL2016 conference: Tourism and cultural landscapes: Towards a sustainable approach, pp. 614-628.

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

Westra, E., 2014. *Building Mapping Applications with QGIS*. Packt Publishing Ltd.

Zhou, Y., Li, N., Wu, W., Wu, J., Shi, P., 2014. *Local spatial and temporal factors influencing population and societal vulnerability to natural disasters*. Risk Analysis: An Official Publication of the Society for Risk Analysis. 34(4). pp. 614-639.