The open discussion version of this paper is available at: Oxoli D, Zurbarán MA, Shaji S, Muthusamy AK. (2016) Hotspot analysis: a first prototype Python plugin enabling exploratory spatial data analysis into QGIS. PeerJ Preprints 4:e2204v4 https://doi.org/10.7287/peerj.preprints.2204v4

Hotspot analysis: a first prototype Python plugin enabling exploratory spatial data analysis into QGIS

Daniele Oxoli[1], Mayra A. Zurbarán[2], Stanly Shaji[3], Arun K. Muthusamy[2]

[1]Politecnico di Milano, Dept. of Civil and Environmental Engineering, Como Campus, Como, Italy
[2]Universidad del Norte, Dept. of Systems Engineering and Computer Science, Barranquilla, Colombia
[3]Politecnico di Milano, Dept. of Electronics, Information and Bioengineering, Milano, Italy

Corresponding author:
Daniele Oxoli[1]

Email address: daniele.oxoli@polimi.it

ABSTRACT

The growing popularity of Free and Open Source (FOSS) GIS software is -without doubts- due to the possibility to build and customize geospatial applications to meet specific requirements for any users. From this point of view, QGIS is one of the most flexible as well as fashionable GIS software environment which enables users to develop powerful geospatial applications using Python. Exploiting this feature, we present here a first prototype plugin for QGIS dedicated to Hotspot analysis, one of the techniques included in the Exploratory Spatial Data Analysis (ESDA). These statistics aim to perform analysis of geospatial data when spatial autocorrelation is not neglectable and they are available inside different Python libraries, but still not integrated within the QGIS core functionalities. The main plugin features, including installation requirements and computational procedures, are described together with an example of the possible applications of the Hotspot analysis.

Keywords: ESDA, Hotspot Analysis, QGIS, Python, FOSS

INTRODUCTION

Exploratory Spatial Data Analysis (ESDA) identifies a collection of 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 years ESDA brought valuable answers to different research fields such as epidemiology, criminology, economy, archaeology, wildlife biology etc. leading to an increasing consideration of these statistical techniques among GIS scientists. This success is mainly due to the specific type of data for which these techniques are intended. Data involved in these analysis has -in fact- a geospatial nature and at the same time describes social, demographic as well as economic attributes. These attributes are intrinsically related to the location to which they refer to and they usually show a strong spatial autocorrelation. According to Cliff & Ord (1973), regression analysis of spatially distributed variables can lead to unreliable statistical inference when proper corrections for spatial effects are not incorporated in the model. This is due to the incorrect assumptions of the independence of observations (Holt, 2007). Conversely, central to

Please cite this paper as: Daniele Oxoli, Mayra A Zurbarán, Stanly Shaji, Arun K Muthusamy (2018) Hotspot analysis: a first prototype Python plugin enabling exploratory spatial data analysis into QGIS. In Marchesini I. & Pierleoni A. (Eds.) Proceedings of the 4th Open Source Geospatial Research and Education Symposium (OGRS2016). Perugia, 12-14 October 2016. https://doi.org/10.30437/ogrs2016_paper_29

ESDA is exactly the spatial autocorrelation in which locational similarity (i.e. observations in spatial proximity) is matched by attribute correlation (Anselin, Sridharan & Gholston, 2007). For these reasons, the use of ESDA is largely accepted as the best practice for the analysis of this kind of data, which nevertheless represents a large share of geospatial information adopted in scientific as well as in socio-economical studies.

Besides the theory, ESDA relies on various software implementations which is a strong triggering factor for spreading the usage of these techniques. One of the most famous ESDA package is included in the proprietary software ArcGIS (*https://www.arcgis.com*), other interesting Free and Open Source Software (FOSS) applications are the Exploratory Spatio-Temporal Analysis Toolkit (ESTAT) (*http://www.geovista.psu.edu/ESTAT*), the Space-Time Analysis of Regional Systems (STARS) (*http://regionalanalysislab.org/index.php/Main/STARS*) as well as the PySAL Python Library (Rey & Anselin, 2010).

In this work we present the implementation of a first prototype plugin to enable the use of some PySAL ESDA tools -i.e. Hotspots analysis with Gi* local statistics (Getis & Ord, 1992)- inside QGIS (*www.qgis.org*); one of the most famous and widely adopted FOSS GIS. The purpose is both, to facilitate the access to this particular type of spatial analysis for users with no advanced programming skills -by exploiting the user-friendly QGIS environment- as well as contributing to the growth of the mapping capabilities of this FOSS GIS software.

PLUGIN DEVELOPMENT

QGIS is not only a GIS software; it is also a geospatial programming environment, which can be used to build geospatial applications using Python. While QGIS itself is written in C++, it includes extensive support for Python programming (Westra, 2014). QGIS Python plugins are based on the Python bindings of Qt framework, which are called PyQt. Specific 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. On the other hand, external Python Libraries need to be imported in order to make them available into QGIS.

The presented plugin is based mainly on PySAL (*http://pysal.github.io*), which allows to compute Gi* statistics. Other required libraries are SciPy (*https://www.scipy.org*), NumPy (*https://www.numpy.org*) and Pyshp (*https://pypi.python.org/pypi/pyshp*). All these libraries are not included in the default QGIS installation and therefore, they need to be installed to make the plugin run. Both documentation about the installation procedure and the source code were made available on GitHub (*https://github.com/stanly3690/HotSpotAnalysis_Plugin*).

PLUGIN FUNCTIONALITIES

Hotspot analysis plugin requires a point shapefile as input having -at least- three correctly assigned attributes which are X, Y projected coordinates -in two separated fields of the attribute table- and a positive numeric attribute. X and Y projected coordinates can be computed by using QGIS field calculator while the positive numeric attribute has to be assigned to any point depending on the user's analysis purpose. Example of this numerical attribute are census data, number of registered crime cases, house prices etc. This information has to be assigned to a pointwise location representative for a parcel of the area under investigation (e.g. city blocks for

cities, municipalities for regions, regions for countries etc.). A common choice is to aggregate events or incidental points (i.e. pointwise information with no associated numerical attribute) by counting the events registered at any user selected parcels. When a multi-polygon shapefile containing this information is available, it is possible to create a point shapefile representative for any parcel (i.e weighted points) by using the Polygon Centroids plugin, included in the Geometry tools of QGIS. At this point, it is possible to run Hotspot analysis plugin by specifying the required attribute fields and a distance threshold, as shown in figure 1. This latter is needed in order to assign Z-scores of the GI* local statistic as well as p-values of the null-hypothesis (i.e. complete spatial randomness) to any point of the dataset by looking at its neighborhoods in a defined region around the point (see e.g. Getis & Ord 1992).

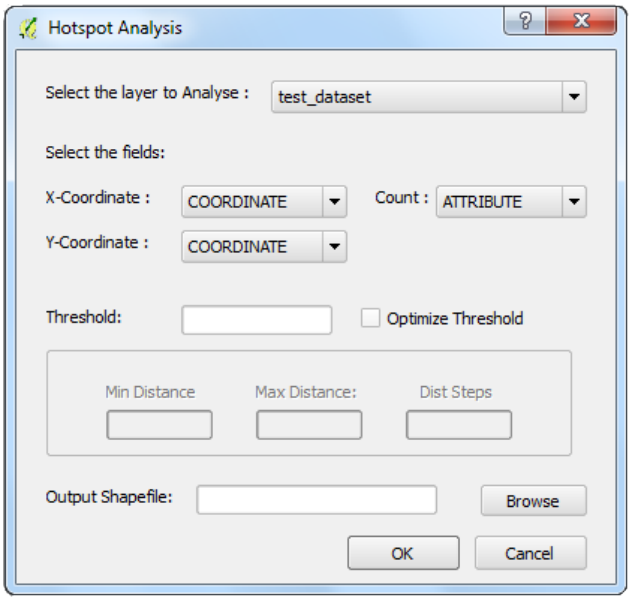

Figure 1. Hotspot analysis plugin interface.

The plugin includes a semi-automatic optimization for the distance threshold selection which allows to estimate the best threshold to adopt based on the Moran's I index (see e.g. Ord & Getis, 1995). Moran's I statistics is included in the PySAL – ESDA tools. The user is asked to specify a distance interval and a distance step with the same units of measure of the projected coordinates system adopted for the input point shapefile. The plugin tests all the distances within the interval by steps and looks when Z-score of Moran's I is maximum. This indicator allows to estimate distance at which the dataset shows higher cluster intensity. The associated distance is used to compute Gi* statistics. The output shapefile contains the three required input attributes in addition to the Gi* Z-score and the p-value, computed at any point location. Combining this two values it is possible to identify if a point is a hotspot or a coldspot. Reference values for Z-score and p-values are associated with the standard normal distribution and the thresholds adopted depend on the specific level of confidence at which the analyst is interested.

An example of hotspot classification is available as a Style Layer Definition (SLD) file for QGIS -together with a test input point shapefile- inside the GitHub plugin repository (*https://github.com/stanly3690/HotSpotAnalysis_Plugin/tree/master/test_data*).

PLUGIN APPLICATION: SENSING ATTRACTIVE LOCATION FOR SLOW-MOBILITY ACTIVITIES USING USER GENERATED CONTENT

An interesting application of Hotspot analysis is the detection of atypical concentration of social media data or user generated content within a region. In this example, the purpose was to identify attractive locations for slow-mobility activities starting from users generated content posted on community Web platforms.

The community platform selected was Wikiloc (*http://www.wikiloc.com*) due to its type of content which are GPX tracks related to outdoor activities (e.g. hiking, biking, running, etc.). This test was carried out with three months of data -from September to December 2015- for the Lombardy Region (northern Italy). GPS waypoints were extracted from the Wikiloc GPX tracks and stored in a PostgreSQL/PostGIS (*http://postgis.net*) database table. Waypoint timestamp was also stored in the database enabling distinctions between waypoints register during weekdays and weekends.

The main focus of the analysis was to understand if the waypoints concentration showed specific spatial patterns inside the study area, in order to identify the most visited locations. One of the most common tools to visualize where a higher density of pointwise data occurs in space is the heatmap. QGIS includes a specific plugin to compute raster heatmaps (*http://tinyurl.com/zlp5chr*) and therefore this analysis option was initially considered and tested. Heatmaps were created by differentiating GPS waypoints registered during weekdays and weekends. Resulting maps are reported in figure 2.

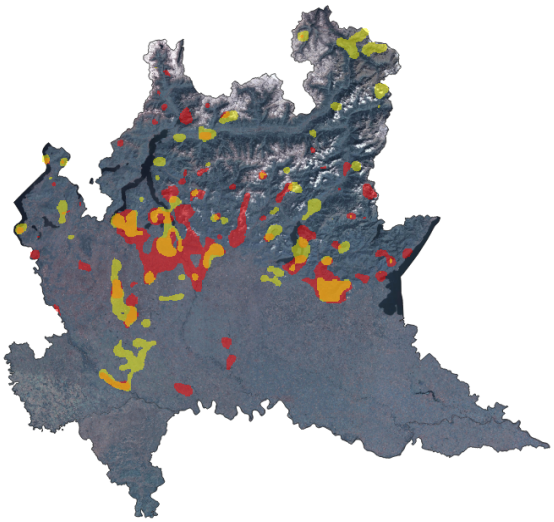

Figure 2. High density clusters extracted from heatmaps of the Wikiloc GPS waypoints registered in weekdays (yellow) and weekend (red). Orange patches represents overlapping clusters. Pixels in the clusters have density values higher than the mean added to the standard deviation from each heatmap.

The main drawback of using heatmaps lies in the fact that both, the type of density function and the visualization parameters -adopted to produce the output map- strongly affect the result. Moreover, density maps such as heatmaps can identify where data clusters exist but not if these are statistically significant.

For these reasons, the Hotspot analysis plugin was involved in order to make less subjective the interpretation of the results. Due to the regional scale at which the analysis aimed to, the territorial parcels selected for data aggregation were the municipalities. The count of waypoints falling into any municipality area was assigned as attribute to the municipalities shapefile (see figure 3). A point shapefile was then created by computing the municipality centroids. Centroids X and Y projected coordinates were also assigned as attributes.

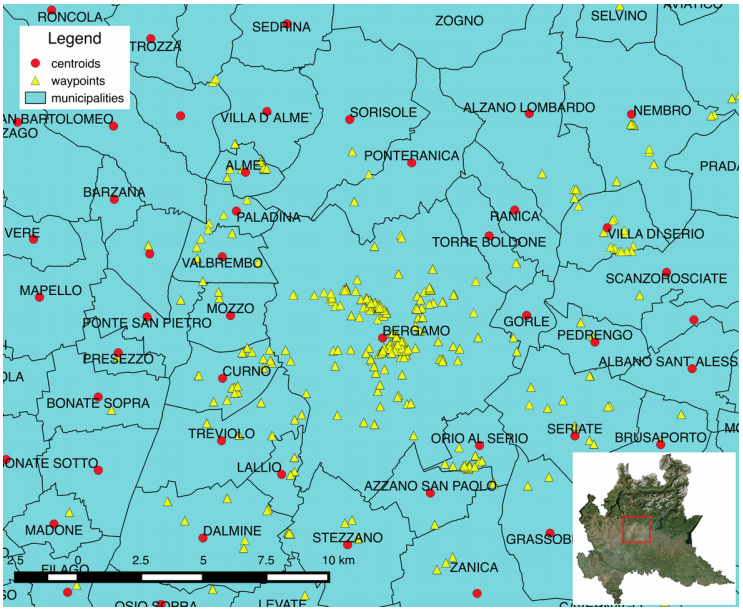

Figure 3. Zoom on some of the GPS waypoints (yellow triangles) scattered over the Lombardy Region municipalities (light blue polygons). The count of waypoints within any municipality is assigned as attribute to the municipality centroids (red dots).

Hotspot analysis was then performed, highlighting important differences between hotspot patterns during weekdays and weekend (see figure 4), as well as identifying most attractive locations (i.e. hotspot clusters) for slow-mobility within the Lombardy Region, according to user's activities (Brovelli, Oxoli & Zurbarán, 2016).

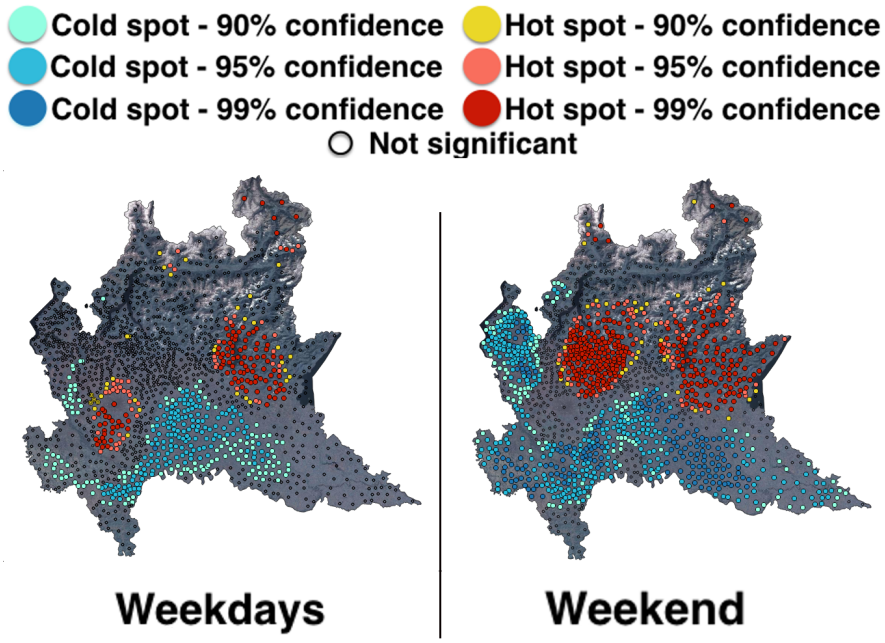

Figure 4. Hotspot analysis applied to the centroids layer for the entire region. Different colors are used to distinguish between hotspots and coldspots according to the computed Z-score and p-value.

By comparing results from heatmaps and Hotspot analysis, it is possible to notice that while the high density cluster patterns follows in general the hotspot patterns, some of the high density clusters lie within not statistically significant or even coldspot locations. This evidence proved the utility of Hotspot analysis for a more robust and precise identification of density clusters and -in turn- of attractive locations. Moreover, the possibility to perform these two different spatial analyses within a unique GIS software was also a valuable factor in order to speed up the analysis process and to enhance critical results comparison, which are key topics from the user's perspective.

CONCLUSION AND FURTHER IMPROVEMENTS

Potential applications of the Hotspot analysis -or more in general of the ESDA- are broad and helpful for manifold scientific fields. The possibility to perform this kind of analysis within QGIS represents a valuable incentive to boost the use of this FOSS GIS among a larger user community. The inclusion of PySAL into QGIS represents a meaningful objective in order to strengthen the capabilities of this software.

The presented plugin, besides being currently a prototype, aims exactly to bring new, fresh and ready-to-use geospatial functionalities to QGIS users. Therefore, further improvements will focus on refining the work done for the Hotspot analysis plugin; first through dependencies reduction by substituting some of the external libraries functionalities with available PyQGIS APIs and then through the inclusion of other ESDA tools from the PySAL core library.

ACKNOWLEDGMENTS

Acknowledgements to the Sustain-T Project (Technologies for Sustainable Development) by Erasmus Mundus for supporting the author and encouraging international cooperation in research.

REFERENCES

Anselin, L. 1999. Interactive techniques and exploratory spatial data analysis. In: P. Longley, M. Goodchild, D. Maguire, and D. Rhind eds. Geographical Information Systems: Principles, Techniques, Management and Applications, John Wiley & Sons, New York, 253–266.

Anselin, L., Sridharan, S., Gholston, S. 2007. Using exploratory spatial data analysis to leverage social indicator databases: the discovery of interesting patterns. *Social Indicators Research*, 82(2), 287-309.

Brovelli, M. A., Oxoli, D., Zurbarán, M. A. 2016. Sensing Slow Mobility and Interesting Locations for Lombardy Region (Italy): a Case Study Using Pointwise Geolocated Open Data. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* 603-607.

Cliff, A., Ord, J. K. 1973. *Spatial autocorrelation*. Pion, London.

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

Holt, J. B. 2007. The topography of poverty in the United States: a spatial analysis using county-level data from the Community Health Status Indicators project. *Preventing chronic disease*, 4(4).

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

Rey, S. J., Anselin, L. 2010. PySAL: A Python library of spatial analytical methods. In: Handbook of applied spatial analysis. Springer Berlin Heidelberg, 175-193.

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