Intrinsic assessment of OpenStreetMap contribution patterns through Exploratory Spatial Data Analysis

Marco Minghini[1,*], Daniele Oxoli[2], Francesco Frassinelli[3] and Maria Antonia Brovelli[2]

[1] European Commission, Joint Research Centre (JRC), Ispra, Italy; marco.minghini@ec.europa.eu
[2] Department of Civil and Environmental Engineering, Politecnico di Milano, Milan, Italy; daniele.oxoli@polimi.it, maria.brovelli@polimi.it
[3] Norsk institutt for naturforskning (NINA), Trondheim, Norway; francesco.frassinelli@nina.no

* Author to whom correspondence should be addressed.

This abstract was accepted to the Academic Track of the State of the Map 2019 Conference in Heidelberg after peer-review.

Compared to traditional geospatial data sources, a major advantage of OpenStreetMap (OSM) is the availability of its full history. In literature, OSM history has been exploited for a number of purposes. The most frequent is intrinsic quality assessment, which – in contrast to extrinsic assessment, where OSM quality is evaluated through comparison against a reference dataset – estimates OSM quality by only looking at its temporal evolution. OSM history has been also explored to gain insights into the project's contribution patterns, e.g. history and profiling of contributors; origin, amount, nature and frequency of edits; spatio-temporal evolution of the whole OSM database – or parts thereof, such as road networks and buildings – in specific areas, or after specific events like natural disasters; and spatial analysis of contributor and contribution patterns.

This work fits into the context of OSM intrinsic assessment by proposing a statistical approach based on Exploratory Spatial Data Analysis, and in particular spatial association [1], aimed at uncovering underlying history-based patterns of OSM data. More in detail, spatial association is investigated in both the univariate and multivariate contexts, i.e. in the cases – respectively – when one variable and multiple variables (together) are examined. The univariate analysis is performed using the Local Moran's I indicator, which provides a robust classification method to detect statistically significant patterns (compared to the hypothesis of randomness) and defines the spatial association type at each location in the dataset [2]. The association type reflects the local characteristics of the variable at each location and its surroundings. Hence it allows detecting clusters, i.e. local patterns of similar (either high or low) values, as well as outliers, i.e. local patterns of dissimilar values (either low values surrounded by high values or viceversa). Instead, the multivariate Geary's c indicator is employed to detect local association patterns resulting from the joint spatial interaction of two or more variables [3]. A multivariate pattern classification comparable to the one of the univariate case is achieved through a novel classification method developed by the authors [4]. This consists of a comparison of local and global centrality measures

Minghini et al. (2019). Intrinsic assessment of OpenStreetMap contribution patterns through Exploratory Spatial Data Analysis
In: Minghini, M., Grinberger, A.Y., Juhász, L., Yeboah, G., Mooney, P. (Eds.). Proceedings of the Academic Track at the State of the Map 2019, 13-14. Heidelberg, Germany, September 21-23, 2019. Available at https://zenodo.org/communities/sotm-2019
DOI: 10.5281/zenodo.3387683

(means and medians) for the computed distribution of the multivariate Geary's c, to produce classification maps of clusters and outliers.

The analysis is performed on Milan Province (Northern Italy), counting a population of more than 3 million inhabitants on a surface of about 1.500 km². This area is sampled using a regular hexagonal grid with side of the hexagon equal to 1000 m, producing a total of 684 cells. The analysis is focused on the history of OSM nodes only, with the following hypotheses: only nodes with at least one tag are considered; a new version of a node is counted only when there is a change in tags (not in geometry); only the nodes which currently exist in the OSM database are considered. With this in mind, for each grid cell a number of history-based variables (mostly derived from literature) are computed: total number of different contributors who have edited OSM nodes; average number of different contributors who have edited each OSM node; average date of creation of the OSM nodes; average date of last edit of the OSM nodes; average number of versions of the OSM nodes; average frequency of update of the OSM nodes. These values are derived from the processing of the OSM Full History Planet file (downloaded in May 2019) and its conversion into a SpatiaLite database after an intersection with the study area, followed by the computation of the variables for each grid cell.

The univariate analysis, performed using the QGIS Hotspot Analysis plugin developed by the authors [5], highlights different spatial associations for the different variables. While some of them (such as total and average number of contributors and average number of versions) clearly show clusters of high values in correspondence of the most urbanized areas and clusters of low values in the non-urban peripheral areas, spatial association patterns are more heterogeneous for other variables such as the average update frequency. Multivariate analyses are then performed to detect the spatial patterns derived from the joint interaction between two and more of the variables considered. Despite each variable has its own spatial pattern when taken alone, their combination (especially when adding more and more variables) highlights not only high and low-value clusters in urban and non-urban areas, but also other interesting clusters and outliers. These unveil peculiar contribution patterns resulting from active local contributors, data imports and mapping parties, and highlight areas where OSM development might need some improvement.

Despite preliminary, the methodology – which, to the authors' knowledge, has been never adopted before in OSM-related research – looks extremely promising to process the complexity of OSM history and transform it into understandable, statistical-based indicators which can shed more light on the intricate phenomenon of OSM local development.

References

[1] Unwin, A., & Unwin, D. (1998). Exploratory spatial data analysis with local statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *47*(3), 415-421.

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

[3] Anselin, L. (2019). A local indicator of multivariate spatial association: extending Geary's C. *Geographical Analysis*, *51*(2), 133-150.

[4] Oxoli, D. (2019). *Exploratory approaches in spatial association analysis: methods, complements, and open GIS tools development*. Doctoral dissertation, Politecnico di Milano, Italy.

[5] Oxoli, D., Prestifilippo, G., Bertocchi, D., & Zurbaràn, M. A. (2017). Enabling spatial autocorrelation mapping in QGIS: The Hotspot Analysis Plugin. *GEAM. Geoingegneria Ambientale e Mineraria*, *151*(2), 45-50.