Data quality analysis tools and their use for the detection of spatio-temporal heterogeneities in VGI contributions

Elias Nasr Naim Elias[1], Fabricio Rosa Amorim[1], Marcio Augusto Reolon Schmidt[1] and Silvana Philippi Camboim[1]

[1]Programa de Pós Graduação em Ciências Geodésicas
Avenida Coronel Francisco Heráclito dos Santos, 210 – Jardim das Américas – Curitiba, Paraná, Brazil
Tel.: +55 41 3361 3153
elias_naim2008@hotmail.com; fabricioamorimeac@hotmail.com; marcio.schmidt@ufu.br; silvanacamboim@gmail.com

Abstract

The quality of data from VGI platforms is a recurring theme in recent research. The spatiotemporal distribution of contributions on such platforms is quite heterogeneous, depending on several factors such as the availability of inputs, the number and motivation of volunteers, among other causes that are still little explored in the literature, especially in developing countries. The present work aims the development of tools to evaluate and visualize quality heterogeneity of the available features on the OpenStreetMap's (OSM) Volunteered Geographic Information (VGI) platform in Brazil. The analyzes were developed using two tools addressing different parameters of data quality. The first one referred to the identification of the history of editions evolution between 2008 and 2020 in different regions from Brazil. The second one focused on the spatialization of the positional discrepancies of the OSM features, in the city of Salvador. In terms of temporal distribution, it was noted that there are not homogeneous patterns, since some regions presented linear growth behaviour. In others, there was a large number of mass contributions. In discrepancy's analysis, it varied from 0.27m to 11.27m, which indicates the heterogeneity of positional data quality. These are the initial results on the development of a toolset to provide users visualization and analysis of these parameters on open platforms.

Keywords: Geospatial Data Quality, Data Heterogeneity, OpenStreetMap.

1 INTRODUCTION

On the steps that permeate the generation of geospatial data, one of the most critical and discussed topics by the scientific community concerns the development of methodologies for the evaluation and representation of the quality of this type of data. It is essential to consider that through time, GIS users began to notice that in most cases, they did not know the quality of the mapped features, which made it difficult to use them in some applications (KOUKOLESTSOS, 2012). In this context, Devillers and Jeansoulin (2006) pointed out the existence of problems related to the evaluation of the quality in geospatial data. They highlighted its intensification from the technological advances and the availability of spatial information in web platforms, beyond the different applications related to the use of spatial data in digital platforms.

About the parameters for evaluating the quality of geospatial data, a series of standards around the world provide the indicators to evaluate the quality of a cartographic product. One of the first standards that approach the quality of geospatial data was the Spatial Data Transfer Standard (SDTS), developed in 1992 and implemented by several governmental and private institutions in The United States (AL-BAKRI, 2012). According to Al-Bakri (2012), the mapping standards around the world that were developed based on the standards established by STDS to define

quality aspects, such as Australian and New Zealand Land Information Council (ANZLIC) and the European Committee Standardization (CEN). The most recent definition is provided by the International Organization for Standardization (ISO) 19.157 (ISO, 2013) in which are approached the indications of positional accuracy, thematic accuracy, completeness, logical consistency and usability.

With the development of technology, it is now possible to obtain spatial data from different sources. The creation of such information, previously restricted to specialists, is now possible through contributions from lay users through collaborative platforms Volunteered Geographic Information (VGI), such as defined by Goodchild (2007). In this context, it is vital to incorporate procedures for quality evaluation, as the diverse sources that contemplate these methodologies (Brovelli et al., 2019). A bibliometric review about the quality of the geospatial data performed by Bielcka and Burek (2019), emphasized that, since 2004, the geospatial data provided by volunteers has become the object of extensive research and these studies analyze, not only the different types of errors but also the behaviour of the volunteers in a particular area.

Data quality on VGI platforms has been the subject of extensive research, being a recurring objective in this field to integrate such data into authoritative geospatial databases. Among these researches, it is possible to quote the research of Jasin and Hamandani (2020), Ibrahim, Ramadan and Hefny (2019), Tian, Zhou and Fu (2019), Minghini and Frassinelli (2019), Brovelli and Zamboni (2018). These works use traditional quality parameters described by the ISO 19.157 (ISO, 2013) as well as additionally the use of parameters intrinsic to VGI data, such as the history of editions, the number of users and collaboration patterns.

Some recent researches that involve the VGI quality parameters have been trying to comprehend how much the characteristics and the dynamics of the contributions interfere in its quality (DROR, DOYTSHER AND DALTOY, 2020, NASIRI et al., 2018). One aspect identified in these works is the heterogeneity of the data; where the quality parameters can vary according to the study area. In this context, for example, in the same area, the features from VGI can be fully mapped in a portion and also have considerable voids in another portion, or the positional precision may vary from 0.10cm to 10m. In developing countries, the number of researches is smaller but, studies like those from Camboim, Bravo and Sluter (2015) indicate an even more explicit heterogeneity since even the official mapping is quite fragmented and dispersed over time. Since these heterogeneities are so present in the VGI data, we propose that if we can analyze and visualize these discrepancies using open source tools, it will be possible to understand their causes better. In this way, we can identify the elements that affect the quality of the OSM data and enable better decision making regarding the use of this information for a given purpose.

2 METHODOLOGY

To work with positional and temporal quality parameters, we divided the analysis into two phases. The first step corresponded to a preliminary evaluation of the behaviour of the history of editions of the OSM's VGI platform in Brazil. The data for the two selected Brazilian urban regions (5x5 km rectangles), from the years 2008 to 2020, come from the University of Heidelberg´s OHSOME Application Programming Interface (API) (https://heigit.org/big-spatial-data-analytics-en/ohsome/). We used an application using the python language, developed by researchers associated with the OSHOME API project, to compute and generate graphs showing the evolution of updates over time.

The second tool developed allows the comparison of data from a reference base with the OSM data. It was developed as a plugin for the QGIS software also using the python language. To demonstrate its application, we used 20 points in the city of Salvador, capital of the state of Bahia. In this analysis, the Euclidean distance between homologous points in the two bases is calculated. These differences were later interpolated on a surface to demonstrate the variation in the positional quality of OSM data in the city.

3 RESULTS

The temporal behaviour of the editions in the study areas analyzed with the first tool was quite distinct. Despite having equal areas and being all urban, one could have more than 100,000 contributions, while the others did not exceed 2,000. Additionally, it was noted that the pattern of contributions over time could vary from area to area. Figure 1, for example, shows two OSM regions that presented different growth patterns. While one of them remained linear throughout time, another presented a mass contribution in 2016. It is essential to consider that these patterns can help to understand better the way in which data were obtained over time in a given location and help to infer the impact of this behaviour on the quality of the available data.

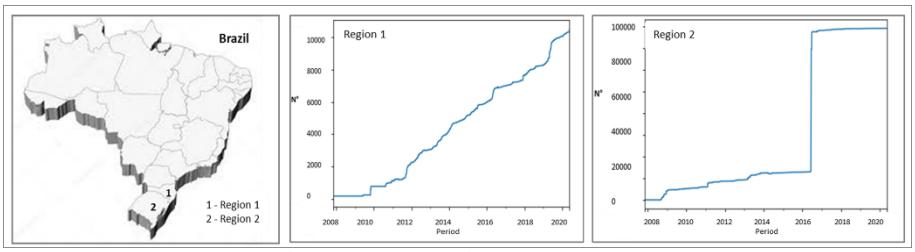

Figure 1. Evolution of OSM contributions from 2008 to 2020 in two areas in Brazil

The graphic differences between regions 1 and 2 reveal discrepancies in terms of the behaviour of the editions. On region 1, the linear growth reflects that the OSM input features were made by gradual contributions of users and regions, while on region 2, the existence of mass contributions caused stability in the number of contributions from 2016.

However, even within these analyzed regions, the data obtained can be heterogeneous. The second tool allows the comparison with official reference bases in order to allow the visualization of positional quality in a region. Figure 2 presents the interpolation of the euclidean distances obtained on the evaluation of the positional accuracy of the OSM's point and linear features of the county of Salvador-BA. We can see that the discrepancies are not evenly distributed throughout the city, but we find regions where the results of positional accuracy are consistently worse. The absolute values in the samples varied from 0.27m up to 11.27m. Tools like this plugin can help users understand the differences in data quality parameters and help advance the understanding of what causes this local inequality.

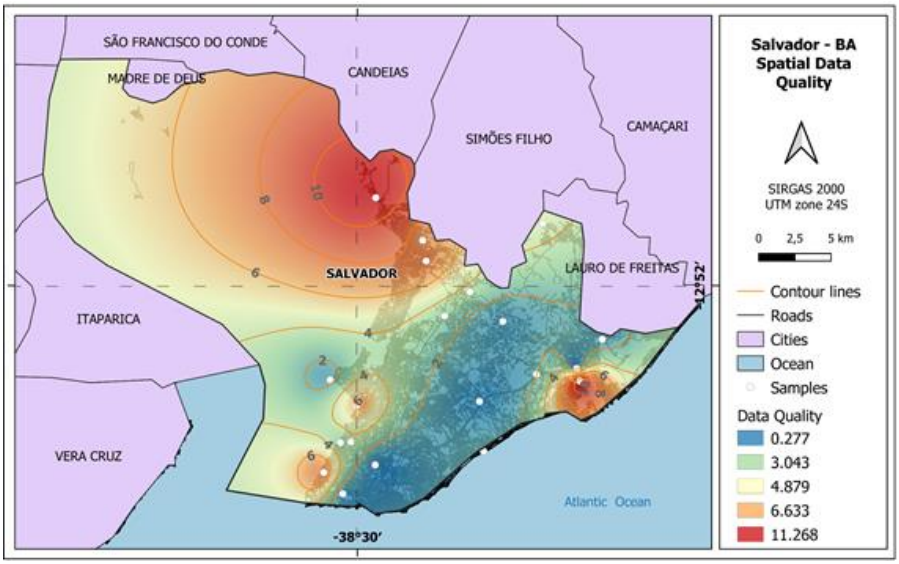

Figure 2. Discrepancies behaviour in Salvador-BA

4 FINAL CONSIDERATIONS

The results of these analyzes corroborate the need for open tools that provide users with knowledge about the spatio-temporal quality of collaborative mapping data. Future studies on the distribution of these parameters can help to model the causes of such heterogeneities. Thus, it is possible to outline actions that contribute to improving the mapping in underrepresented areas. Additionally, by knowing more about the nature of the distribution of contributions, it is possible to plan the integration of these data with official maps and contribute to mapping in Brazil that, due to its extension and lack of investment in cartography, require updated geospatial data.

References

Al-Bakri. Developing Tools and Models for Evaluating Geospatial Data Integration of Official and VGI Data Sources. 2012. PhD. School of Civil Engineering and Geosciences, Newcastle University.

Bielecka, E., Burek, E. (2019). Spatial data quality and uncertainty publication patterns and trends by bibliometric analysis. Open Geosciences, 11(1), pp. 219-235.

Brovelli, M. A.; Boccardo, P.; Bordogna, G.; Pepe, A.; Crespi, M.; Munafò, M.; Pirotti, F. Urban Geo Big Data. 2019. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, FOSS4G 2019* - Academic Track. Bucharest: Romania.

Brovelli, M. A., Zamboni, G. 2018. A new method for the assessment of spatial accuracy and completeness of OpenStreetMap building footprints. *ISPRS International Journal of GeoInformation*, 7(8), pp. 1-25.

Camboim, S.P.; Bravo, J.V.M.; Sluter, C.R. 2015. An Investigation into the Completeness of, and the Updates to, OpenStreetMap Data in a Heterogeneous Area in Brazil. *ISPRS International Journal of GeoInformation*, 4, pp 1366-1388.

Devillers, R., Jeansoulin, R. 2006. Spatial Data Quality: Concepts. In *Fundamentals of Spatial Data Quality* R. Devillers and R. Jeansoulin, (EDS). Ch. 2, pp.31-42. London: ISTE Ltd.

Dror T., Doytsher Y., Dalyot S. 2020. Investigating the Use of Historical Node Location Data as a Source to Improve OpenStreetMap Position Quality. In *Open Source Geospatial Science for Urban Studies* A. Mobasheri (EDS). Lecture Notes in Intelligent Transportation and Infrastructure. Springer, Cham.

Goodchild, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp. 211-221.

Ibrahim, M. H., Darwish, N. R., Hefny, H. A. 2019. An Approach to Control the Positional Accuracy of Point Features in Volunteered Geographic Information Systems. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(6), pp. 169-175.

ISO 19157. 2013. Geographic Information - Data Quality. *International Organization for Standarization*.

Jasim, S., Al-Hamadani, O. 2020. Positional Accuracy Assessment for Updating Authoritative Geospatial Datasets Based on Open Source Data and Remotely Sensed Images. *Journal of Engineering*, 26(2) pp. 70-84.

Koukoletsos, T. 2012. *A Framework for Quality Evaluation of VGI linear datasets*. PhD., University College London.

Minghini, M., Frassinelli, F. 2019. OpenStreetMap history for intrinsic quality assessment: Is OSM up-to-date? *Open Geospatial Data, Software and Standards*, 4(1), pp. 9.

Nasiri, A.; Ali Abbaspour, R.; Chehreghan, A. Jokar Arsanjani, J. 2018. Improving the quality of citizen contributed geodata through their historical contributions: The case of the road network in OpenStreetMap. *ISPRS International Journal of GeoInformation*, 7(7), pp. 253.

Neis, P., Zipf, A. 2012. Analyzing the contributor activity of a volunteered geographic information project—The case of OpenStreetMap. *ISPRS International Journal of GeoInformation*, 1(2), pp. 146-165.

Tian, Y., Zhou, Q., Fu, X. 2019. An Analysis of the Evolution, Completeness and Spatial Patterns of OpenStreetMap Building Data in China. *ISPRS International Journal of GeoInformation*, 8(1), pp. 35.