DOI 10.1590/s1982-21702023000200005

ORIGINAL ARTICLE

Exploring spatio-temporal patterns of OpenStreetMap (OSM) contributions in heterogeneous urban areas

Elias Nasr Naim Elias[1,2] - ORCID: 0000-0003-2289-5055

Fabricio Rosa Amorim[1] - ORCID: 0000-0002-6670-2131

Marcio Augusto Reolon Schmidt[1] - ORCID: 0000-0003-2716-2360

Silvana Philippi Camboim[1] - ORCID: 0000-0003-3557-5341

[1]Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas, Curitiba - Paraná, Brasil.
E-mail: elias.naim@eng.uerj.br, fabricioamorimeac@hotmail.com, marcio.schmidt@ufu.br,
silvanacamboim@gmail.com

[2]Universidade do Estado do Rio de Janeiro, Departamento de Engenharia Cartográfica,
Rio de Janeiro - Rio de Janeiro, Brasil.
E-mail: elias.naim@eng.uerj.br

Received in 01st November 2022.
Accepted in 30th March 2023.

Abstract:

The potential of intrinsic parameters to estimate geospatial data quality on Voluntary Geographic Information (VGI) platforms is a recurrent theme in Cartography. The spatial-temporal distribution in these platforms is very heterogeneous, depending on several factors such as input availability, number, and motivation of volunteers, especially in developing countries. The most recent approaches have been aiming to detail temporal patterns as an additional measure of quality in VGI. This research proposes a methodology to identify and analyze the behavior of the contribution parameters over time (2007-2022) of the OSM platform and differentiates the influences that affect its growth. Part of the Metropolitan region of Curitiba was the study area, subdivided into 1 x 1 km cells. The cumulative growth of contributions was calculated and later adjusted using a Logistic Regression. The obtained parameters made it possible to identify abruptly growing cells caused by external data import, mass contributions, or collective mapping activities. In addition, heterogeneity in the growth of the data available in OSM over time was evident. Furthermore, the proposed methodology promoted the investigation of a new indicator of intrinsic quality based on modelling the spatiotemporal evolution of OSM feature insertions.

Keywords: Geospatial Data Quality; Collaborative Mapping; Intrinsic Parameters; Logistic Regression.

How to cite this article: Elias ENN, Amorim FR, Schmidt MAR, Camboim SP. Exploring spatio-temporal patterns of OpenStreetMap (OSM) contributions in heterogeneous urban areas. *Bulletin of Geodetic Sciences*. 29(2): e2023005, 2023.

1. Introduction

OpenStreetMap (OSM) is one of the world's largest and most popular platforms for collaborative collection (Teimoory, Abbaspour and Chehreghan 2021; Brovelli and Zamboni 2018), with more than 7.8 million registered users, overtaking 16.7 billion edited features. Considering the promising aspects of VGI, focusing on OSM, different researchers worldwide have focused efforts on comprehending its characteristics, especially in terms of data quality and viability in integrating processes (Brovelli and Zamboni 2018; Nasiri et al. 2018; Zhang, Malczewski 2017; Sehra, Singh and Rai 2017; Haklay 2010;).

Geospatial data quality is essential for topographic mapping and is part of the specific standards for its production (ISO 19157 2013; DSG 2015). In VGI, this measurement is even more indispensable because the contributed data are heterogeneous and can be affected by several factors related to the area, profile of the contributors, quantity, and dynamics of the contributions. Extrinsic approaches from comparisons with topographic mapping (Brovelli and Zamboni 2018; Zhang and Malczewski 2017; Haklay 2010) and intrinsic (Sehra, Singh and Rai 2017) characterised by analysis of edition history, the number of collaborators and contributions describe the procedures for evaluating VGI's quality.

The evaluation of intrinsic parameters is even more relevant in developing countries, where the chronic lack of resources for cartography often results in the absence of updated data for comparison. In addition, this situation makes the data from VGI even more necessary to complement existing topographic mapping (Camboim, Bravo and Sluter 2015). In Brazil, for example, research by Silva and Camboim (2020) reveals that on the 1:25.000 scale, little more than 5% of the country's extension has available topographic mapping. Thus, understanding the data's quality in these regions is essential.

The classic intrinsic parameters, characterized by the number of contributions, contributors, and edition history, were based on Linus Law. This law applied to the VGI context, as addressed by Haklay et al. (2010), states that as the quantity of contributors in a given spatial unit increases, the greater and better known its quality. In the early days of intrinsic parameter research, the only temporal variable was the date of the last data edition. However, Paiva and Camboim (2021) demonstrated that this approach alone would not explain the entire behaviour of quality in VGI. Such aspect converges with research recently conducted aimed at detailing temporal patterns and thereby improving the understanding of the spatial distribution of its quality (Le Guilcher, Olteanu-Raimond and Balde, 2022; Grinberger et al., 2021; Brückner et al., 2021; Witt, Loos and Zipf 2021; Arsanjani et al. 2015; Gröching, Brunauer and Rehr 2014).

Concerning mathematical modelling, Grinberger et al. (2021) measured the number of contributions over time to detect events on the OSM platform worldwide. The authors proposed procedures to identify these large-scale data production events in the history if OSM and analyze their patterns. Brückner et al. (2021) estimated the completeness of OSM retail stores in Germany. Both studies are based on accumulated contributions over time and use a four-parameter logistic regression model in their data. This model is described from an "S" shaped curve (sigmoid function). This curve format can be associated with the pattern of contributions in a given area, which starts with few contributions, increases and then gradually stabilizes over time. Brückner et al. (2021) used other regression models in their analysis. When presenting the logistic function to estimate the behavior of the OSM data, Grinberger et al. (2021) related the characteristics of the curve with the premises of the work developed by Gröchenig, Brunauer, and Rehr (2014), who proposed an approach to identify regional and temporally different developments associated with mapping evolution. The proposed model allowed us to classify the different stages of activity in the contributions accomplished in OSM, such as Start, Growth, and Saturation. In this context, Arsanjani et al. (2015) proposed a Contribution Index (CI) based on an analysis the spatio-temporal patterns by OSM contributions. CI encompasses the number of contributions, average number of versions, average number of attributes, and number of users. In addition to the research mentioned above, Witt, Loos and Zipf (2021) analysed the impact of significant imports in OSM for the Netherlands and India, and Guilcher, Olteanu-Raimond and Balde (2022) studied the evolution of massive imports in OSM in France.

Considering the questions presented, as there are a finite number of features in a region, this research aims to study the evolution of contributions in a homogeneous cell, looking for patterns that vary according to the location of this cell, working explicitly with data in large urban centers of Brazil. The hypothesis is that if the existing patterns are known, it will be possible to determine in the future which stage a particular cell is and, therefore, how close it is to having its complete mapping. This information is beneficial for using more robust data and encouraging mapping in regions that are still poorly mapped, thereby reducing heterogeneities, especially in the poorest and peripheral areas.

This research proposes a methodology to identify and analyze the behavior of spatio-temporal parameters of the OSM contributions and differentiate the influences that affect the accumulated growth of the insertion of features over time, either by mass contributions or mapathons. In addition, the interaction between official and collaborative data and the synergy obtained by importing data into OSM were analyzed. Finally, the adequacy of the logistical model was verified for the accumulated count of point, linear, and polygonal features mapped in a specific area, and the measure of the smallest possible size of the surrounding rectangle. The study was conducted in an area in the metropolitan region of Curitiba (Brazil). In this context, it was possible to identify and compare different influences that affect a big metropolis and validate them to its adjacent regions.

1.1 A current panorama of geospatial big data integration and analysis

The current paradigm of the technological environment constitutes an era of Big Data, with all the unique opportunities it presents for the use and production of spatial data (Robinson et al., 2017). The territory portraying, formerly strongly restricted to official agents' and traditional methods, costly, and limited to specialists, nowadays happens dynamically by several agents, including the citizens themselves, through several IT tools. Globally, proposals for alternative sources to traditional ones and the reuse of open geospatial or statistical data have been discussed. The United Nations, through the Statistics Division (UNSD) and the Committee of Experts on Global Geospatial Information Management (UN-GGIM), emphasizes the importance of this vision in documents such as the United Nations Integrated Geospatial Information Framework - UN-IGIF (UN-GGIM, 2022). Countries such as Australia are at the forefront of this integration (Kitchin, 2015). The trend of the general use of Big Data, including diverse sources such as social networks, sensors, and device monitoring, among others, is to improve official databases (Tam and Van Halderen, 2020). In Brazil, large government databases, such as RAIS - Annual Report of Social Information (Ministry of Labour and Previdence, 2022) and CNEFE - National Register of Addresses for Statistical Purposes (IBGE,2022), and cadasters in various instances have the potential to be leveraged. The advantage is to portray specific scenarios with comprehensiveness and dynamism, as in the case of RAIS, created to manage labor data, and which can be important information about companies, providing addresses of Points of Interest that can be geocoded and integrated with existing databases. For instance, Paiva and Camboim (2021) compared data on economic activity permits available at the Open Data Portal of Curitiba City Hall with the data quality parameters from OpenStreetMap.

There is also research on recent applications in urban areas using open geospatial data, whether they are mined from proprietary applications based on crowdsourcing, government open data, and SDis or involve collaborative mapping. For example, Cerqueira and Diniz (2022) explored the distribution of urban equipment to identify primary and secondary centralities in the Metropolitan Region of Belo Horizonte, Minas Gerais. The authors used Google Places of Interest data. In addition, Paiva and Camboim (2022) worked with official and collaborative data from the capital city of Minas Gerais to search for intrinsic geospatial data quality models. Finally, Elias et al.

(2020) explored the extrinsic quality of road axes from the OSM's VGI platform in Salvador, Bahia, compared with the authoritative municipal dataset. The parameters analyzed were positional accuracy, thematic accuracy, and completeness, exposing data heterogeneity in the region.

2. Methodology

The first step in obtaining and analyzing the spatio-temporal patterns of OSM contributions was to subdivide the study region into a grid with cells of 1x1 km (considering approximately, the equivalence along the Equator line). This procedure is essential for identifying the micro patterns of OSM collaboration. The second step corresponds to obtaining the monthly and daily accumulated amount of point, linear and polygonal features inserted into the platform, from 11/2007 to 10/2022, for each evaluated cell. The third step included data adjustment from logistic regression and extraction of parameters for each cell. The fourth step was to create visual representations of the parameters to observe the spatial aspects influencing the contribution patterns.

This study area covers a part of the NUC of the Metropolitan Region of Curitiba (Região Metropolitana de Curitiba, RMC), located in southern Brazil. As described by COMEC (2022), the RMC has 3.223.836 inhabitants, Brazil's eighth most populous metropolitan region. In addition, it is the second-largest metropolitan region in the country, with an extension of 16.581,21 km^2. Figure 1 presents a location map of the study area.

An important aspect worth mentioning concerns the heterogeneous nature of municipalities. For example, in the land use map (Figure 1), Curitiba concentrates most of its extension as built-up areas, whereas the others present some parts with vegetation and exposed soil. This characteristic influenced cell delimitation in the analyses, prioritizing urbanized areas. Selection is necessary because there is a direct relationship between the amount of existing information, local population, and number of contributors and contributions. However, although the cells are primarily distributed in built-up areas, the contribution dynamics differ between the metropolis core and peripheral areas.

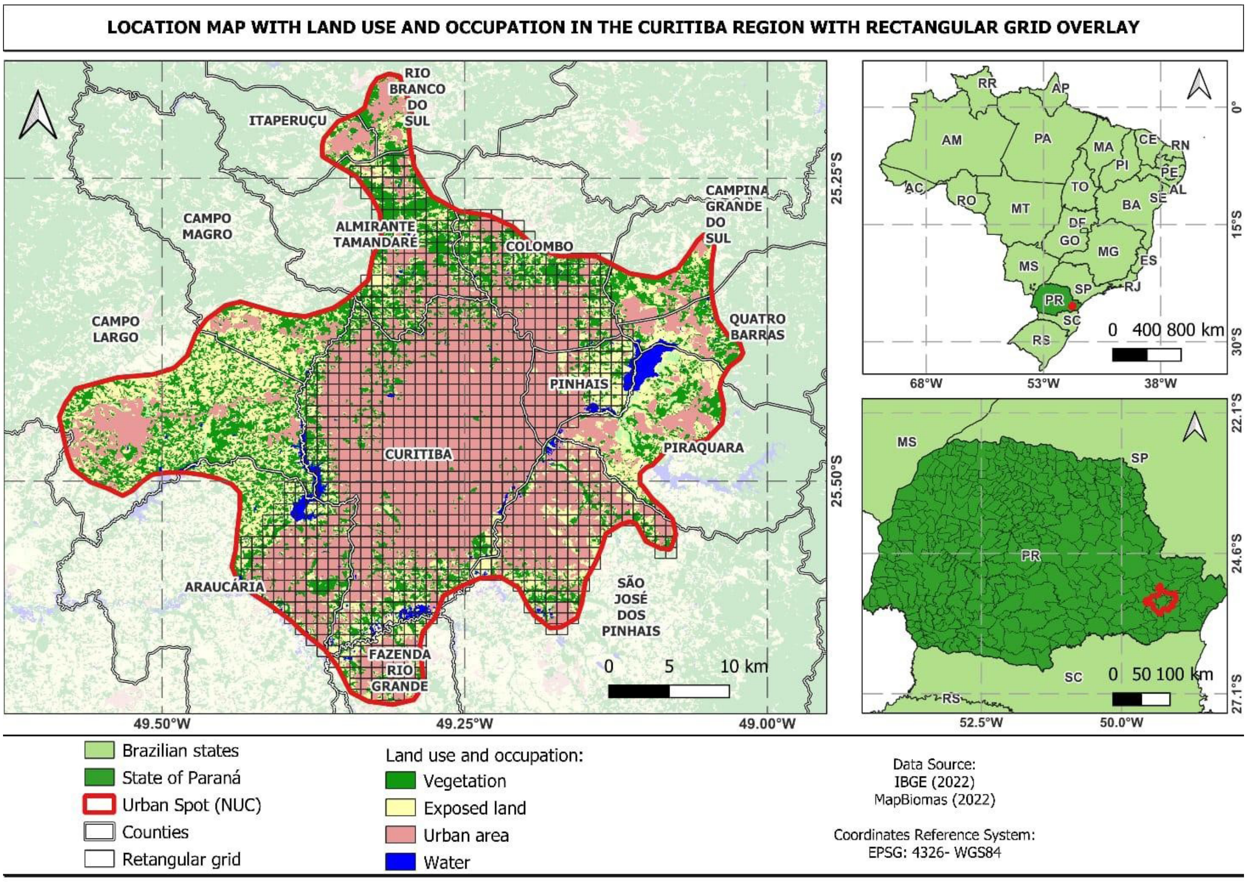


Figure 1: Localisation map of the study area.

2.1 Obtaining the data and adjustment of the logistic curve

The OHSOME Application Programming Interface (API), at https://api.ohsome.org, was used to obtain the number of features inserted over time into OSM. This API allows the extraction of the history of feature editions in OSM, the number of contributors in a specific area or set of features, adding up the information, and even obtaining the set from specific feature categories. Python was used as the programming language was Python, and software QGIS 3.x was used for data gathering and manipulation. The API documentation makes explicit that data extraction must be accomplished through specific queries by delimiting the geographic coordinates in the region of interest. Then, an interactive process that takes the cell's bounding box as input calculates the contribution data and performs mathematical modelling for each. From the contribution history accumulated over time, data modelling was accomplished through a four-parameter logistic regression, as shown in Equation 1.

$$y = a + \left(\frac{b-a}{1+ e^{(c-x)/d}}\right) \tag{1}$$

In Equation 1, parameters a, b, c, and d model the curve and represent the upper asymptote, lower asymptote, midpoint of the logistic curve (on the x-axis), and steepness of the logistic curve. The x value refers to the normalized monthly value, and y represents the accumulated number of features inserted in OSM. Figure 2 shows the logistic curve parameters and zones.

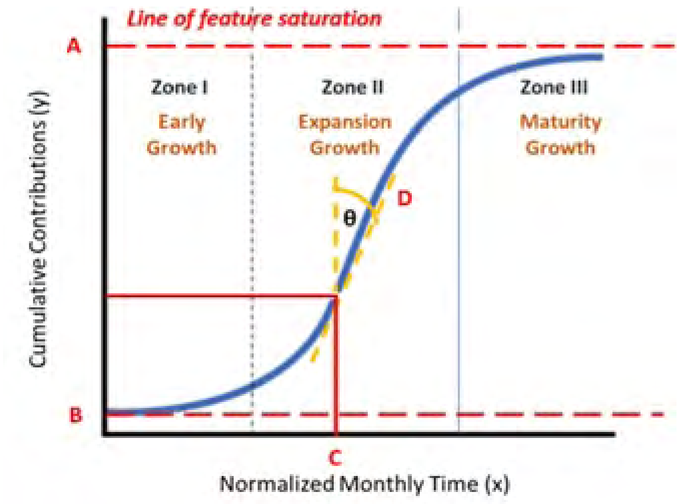

Line of feature saturation
A
B
C
D
θ
Zone I
Early Growth
Zone II
Expansion Growth
Zone III
Maturity Growth
Cumulative Contributions (y)
Normalized Monthly Time (x)

Figure 2: The Logistic Curve parameters and zones.

As shown in Figure 2, the curvature parameters describe its characteristics, from which conclusions can be derived from the contribution patterns over time. For example, in Brückner et al. (2021), one of the criteria used was the value of the upper asymptote relative to the most recent number of features. Furthermore, considering that parameter d is related to the slope of the curve, low values in the regression may indicate specific aspects of the accumulated growth pattern. For instance, in regressions that present abrupt leaps, parameter c tends to be like the month (value x) in which the largest insertion of features occurred.

Figure 2 presents the relationship between the behavior of the curve and the different stages of contributions from zones I, II, and III. Zone I characterizes the initial stage of the mapping activities. Zone II represents the growth of the number of contributions, and zone III characterises a saturation stage in which a given area reaches the maximum limit of features that can be inserted. An important variable analyzed in this research was the presence of high slopes in zone II of the curves. When this slope is too high, the θ angle tends to zero, and it is possible to detect the model areas with significant contributions or data import. The term "cell with abrupt growth" was used in this study.

In addition to the regression graphs, it is possible to obtain the number of monthly inserts in this step. This procedure calculates the number of available features in a specific month subtracted from the monthly value that precedes it. This analysis showed the relationship between the number of features in the highest insertion month and the most recent date. Finally, the regressions were analyzed similar to the methodology presented by Grinberger et al. (2021) by calculating the Normalized Root Mean Squared Error (NRSME) of the median value of each cell's features.

In obtaining and modelling data, it was noticed that Curitiba presented an atypical behavior in the contribution pattern compared to the other municipalities that are part of the NUC. The analysis of monthly growth and obtained values showed this aspect. Furthermore, abrupt jumps can be observed in the number of features inserted over time, in addition to the organic growth (characterized by gradual contributions and different collaborators). Considering that Curitiba is the largest municipality in the evaluated region, it is worth highlighting that the Institute of Research and Urban Planning of Curitiba (Instituto de Pesquisa e Planejamento Urbano de Curitiba - IPPUC) maintains an open geospatial data platform. This website provides points with parcel addresses, urban equipment, and street axes that do not occur in smaller municipalities. It was then identified that the insertions of the data made available by IPPUC into the OSM appeared as abrupt leaps in the curve. For example, Figure 3 presents the arrangement of the features in the OSM considering the presence (A) and absence (B) of the IPPUC data.

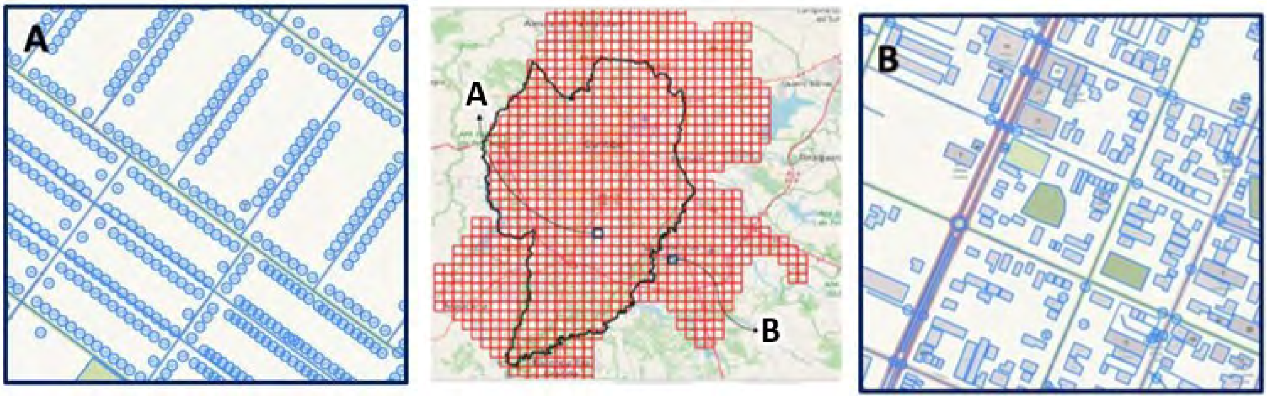


Figure 3: Arrangement of OSM features in Curitiba (A) and São José dos Pinhais (B) cells.

As shown in region A of Figure 3, it is possible to observe the set of point features around the blocks. Such elements correspond to the address information of each parcel available in the region. IPPUC (2022) provides, in shapefile format, more than 300.000 parcels in the municipality with their respective addresses available. It is worth noting that Brazilian municipalities do not commonly make such data available, including those belonging to the NUC. For example, in region B of Figure 3, there were no address points.

All point features with the tag key "addr:street" were filtered. The analysis produced graphs with the model and parameters of the logistic regression, the graph of the monthly contributions, a list of NRMSE values, the quotient between the number of features in the highest contribution month and the most recent date, and the relationship between the number of point, linear and polygonal features. In addition to the described information, cells that presented θ = 0° at some point of curve generation were also computed. These cells are essential because they allow identifying different contribution patterns over time.

2.2 Applications development

The main Python libraries used in the application were spicy to optimize, matplotlib to calculate the logistic regression and plot the data, and NumPy for mathematical operations. The algorithm takes as inputs both the contribution files and the bounding boxes of the cells, generating the output lists after the computation. Figure 4 presents a flowchart of the procedure. Developing an interface enables the user to select the target region, insert data, and calculate the regression, thereby supporting data extraction from different Brazilian regions. Figure 5 shows the interface. Because the contribution patterns tend to present heterogeneous characteristics according to the region, this software allows the selection of the target area and time interval.

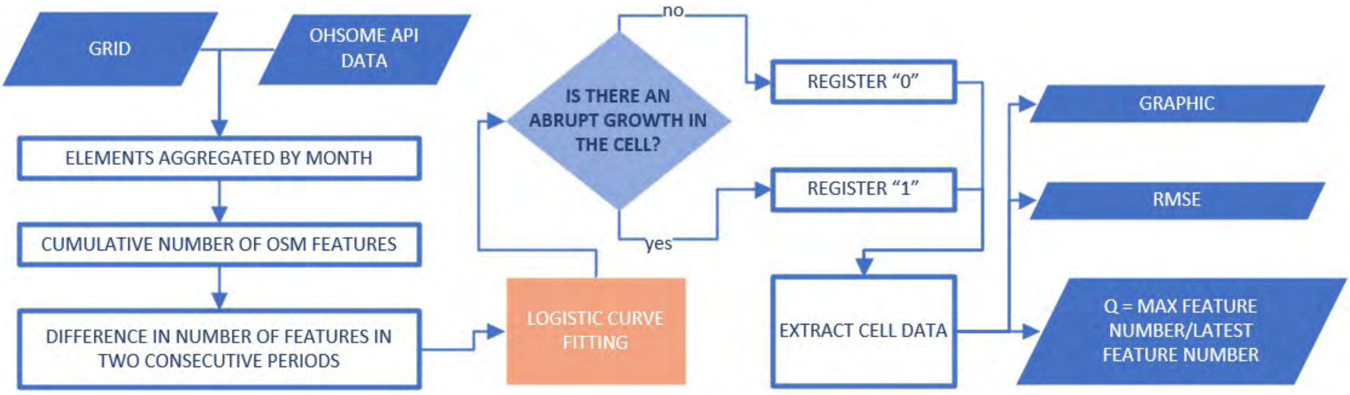
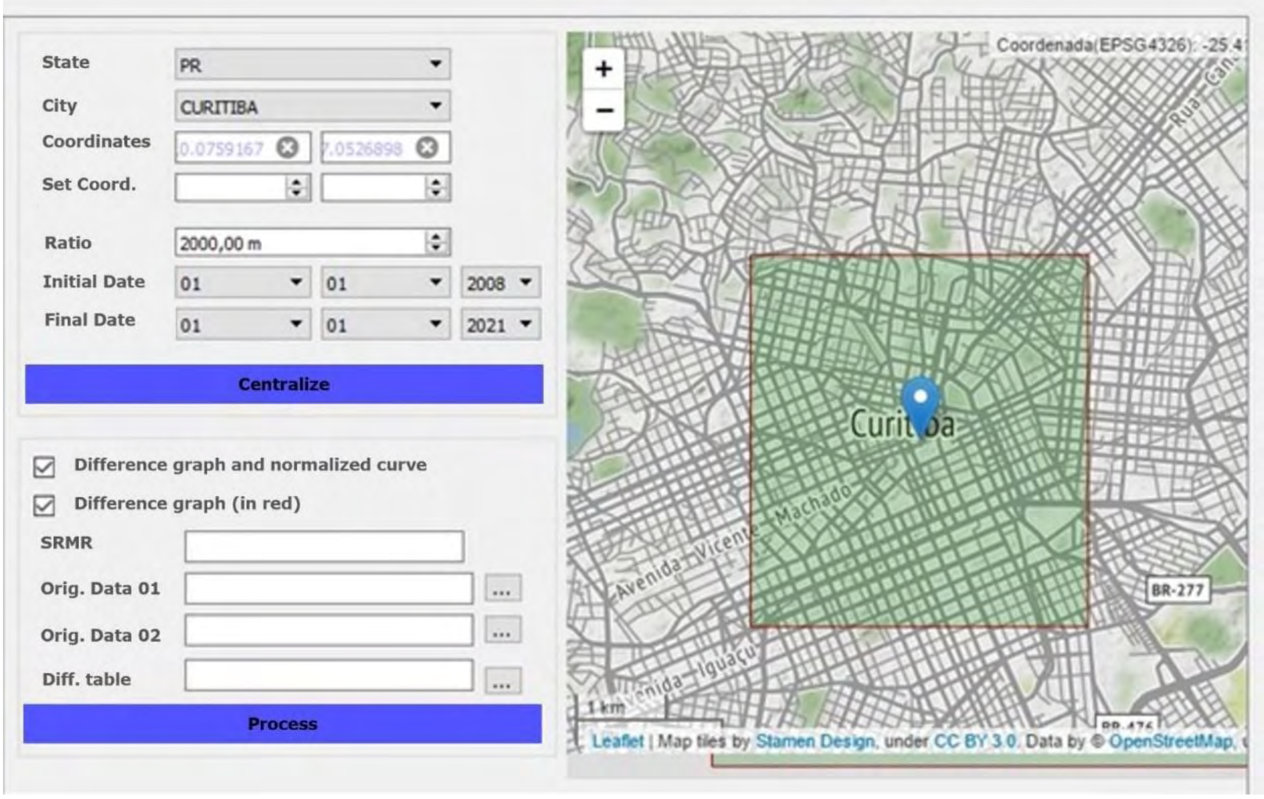

GRID
OHSOME API DATA
ELEMENTS AGGREGATED BY MONTH
CUMULATIVE NUMBER OF OSM FEATURES
DIFFERENCE IN NUMBER OF FEATURES IN TWO CONSECUTIVE PERIODS
LOGISTIC CURVE FITTING
IS THERE AN ABRUPT GROWTH IN THE CELL?
no
yes
REGISTER "0"
REGISTER "1"
EXTRACT CELL DATA
GRAPHIC
RMSE
Q = MAX FEATURE NUMBER/LATEST FEATURE NUMBER

Figure 4: Sequence for extracting the information.



State: PR

City: CURITIBA

Coordinates: 0.0759167 | 7.0526898

Set Coord.:

Ratio: 2000,00 m

Initial Date: 01 | 01 | 2008

Final Date: 01 | 01 | 2021

☑ Difference graph and normalized curve

☑ Difference graph (in red)

SRMR:

Orig. Data 01:

Orig. Data 02:

Diff. table:

Centralize
...
Process
+
−
Coordenada(EPSG4326): -25.4
Rua Gan
Curitiba
Avenida Vicente Machado
Avenida Iguaçu
BR-277
BR-476
1 km
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap

Figure 5: Software Interface for computing the regression.

Considering the interface in Figure 5, the user can select the target municipality and insert the geographic coordinates of the region of interest and the length of the enclosing rectangle from which the data will be obtained. Finally, the target interval and path to store the lists with the results were selected. The interface icons were derived from PyQT and developed using qt designer, in addition folium, for visualizing the base map. The developed scripts are available for downloading in the GitHub online repository.

3. Results and Discussion

Based on the described methodology, 1074 cells with 1x1km had their behaviour evaluated. Figures 6 and 7 show the representation of the existing number of features in OSM and the quotient between the number of points features and the total number of features.

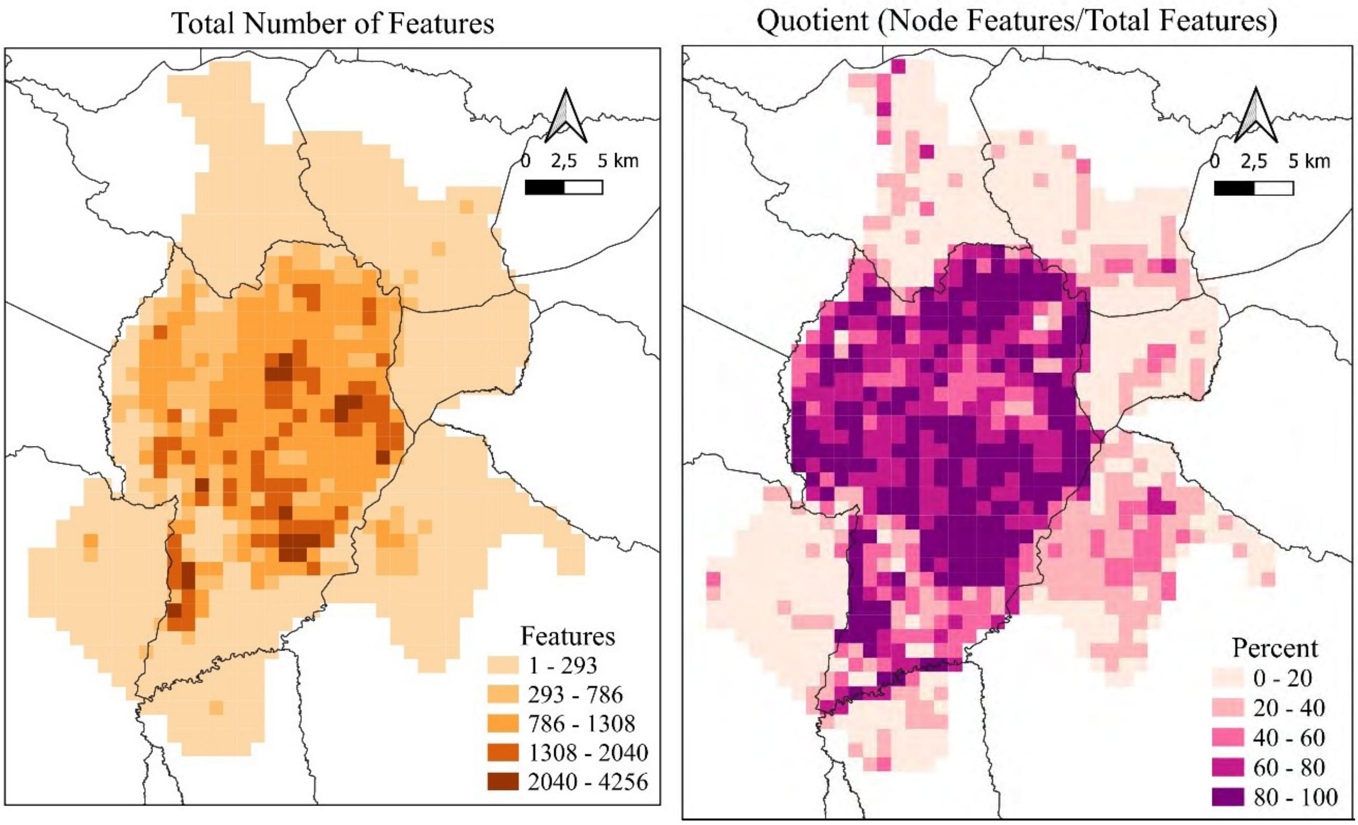


Figure 6A Figure 6B

Figure 6: Features available in OSM (Figure 6A) and the percentage of point features (Figure 6B).

Concerning data availability, areas close to the central region of Curitiba contributed more significantly. The mean and median numbers of features in Curitiba were 804.59 and 834. In the other regions, these values were 92.63 and 65 features, respectively. In addition, it was possible to notice a convergence with the population data, considering that Curitiba is equivalent to approximately 2/3 of the NUC region, in addition to the higher rate of urbanization of the capital. It was also noticed that the central region of the municipality had a higher number of cells, with a predominance of points, more than 60% of the total (Figure 6B). In the regression analysis procedures, in some cells, the value of the θ angle was 0° (d parameter), predominantly in regions with abrupt leaps in accumulated contributions. In data processing, these cells were identified from an overflow message directly related to the high percentage of feature inserts in a month about the total accumulated. The insertion of points into Curitiba's addresses justifies this aspect. The maps in Figures 7A and 7B show the spatialization of cells with abrupt growth and the calculated percentage.

When analyzing Figures 7A and 7B, an initial statement is that the imports of addresses influence the whole pattern of contributions of Curitiba since a single month was equivalent to more than 80% of the total features. Other municipalities of the NUC did not show this aspect, and the percentages varied predominantly between 0% and 40%. Of the 527 cells that intercepted Curitiba, in 233, the percentage (Figure 7B) was greater than or equal to 70%, and abrupt jumps accounted for more than 80% of them (Figure 7A). Overflow messages were captured in all cells in which the percentage was higher than 80%. The abrupt growth was not predominant in the cells of other municipalities of the NUC, considering that regions with low contribution characterized those that occurred. An exception to this behavior occurred in a cell in Araucária, where abrupt growth indicated a significant contribution in a residential area. In this context, despite being predominantly urban, as presented in the land-use map in Figure 1, the neighboring municipalities of the NUC include agricultural and forestry areas, which naturally have a different pattern of features than urban areas. Figure 8 presents examples of the behavior of the curve obtained in the logistic regression of the evaluated cells in different aspects.

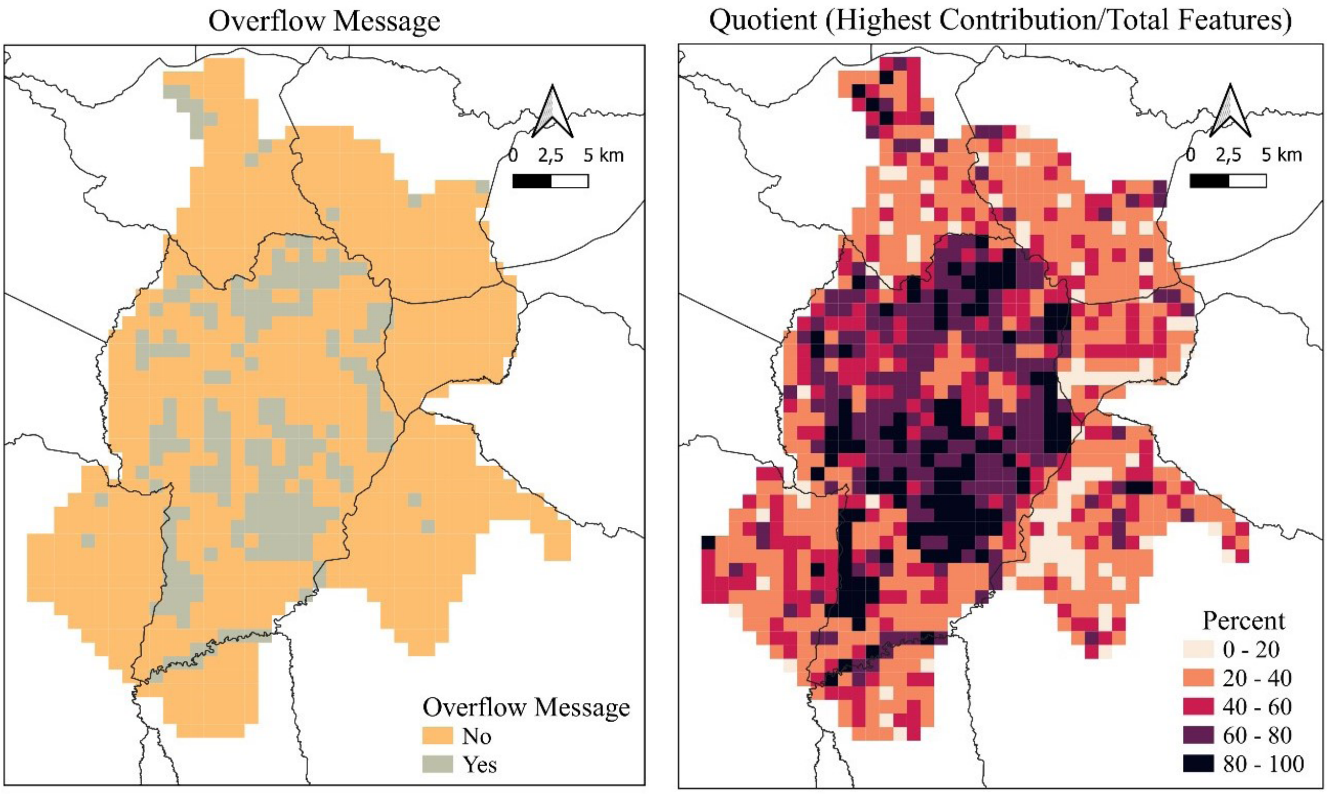


Figure 7A

Figure 7B

Figure 7: Cells with overflow (Figure 7A) and % of the largest contribution month (Figure 7B).

In the graphs in Figure 8, blue represents the accumulated growth of contributions and red represents the logistic regression curve. The jump in the number of inserted features in graph A is due to the insertion of IPPUC address points. There were 1248 contributions on a single day, corresponding to 80% of the total. In graph B, the inserts are evident, as 299 features were inserted in a single month, corresponding to 39% of the total. In turn, this importation occurred in a mix of organic features with proportional amounts, which generated a smoother curve. It is possible to describe the smoothness of the curve in terms of the slope d parameter (Equation 1), where a lower the value indicates a steeper slope. The accumulated contributions in graphs C and D do not show any leaps. However, in graph C, it is possible to notice a mix of significant contributions in short periods, beyond the fact that Zone III of the curve is close to reaching a saturation level, considering the last year's stability. In this region, most building polygons were mapped. In these features, the points addresses may not have been inserted. Instead, the editions may have included the insertion of address tags into an existing building. This justifies the absence of abrupt jumps in this cell.

The import of addresses into Curitiba occurred between 2018 and 2019. Therefore, creating graphs of the accumulated monthly quantity of features from 01/01/2018 to 01/01/2020 enabled identifying and evaluating the contribution patterns and their magnitudes, in addition to the abrupt jump behavior. Figure 9 shows examples of the obtained results.

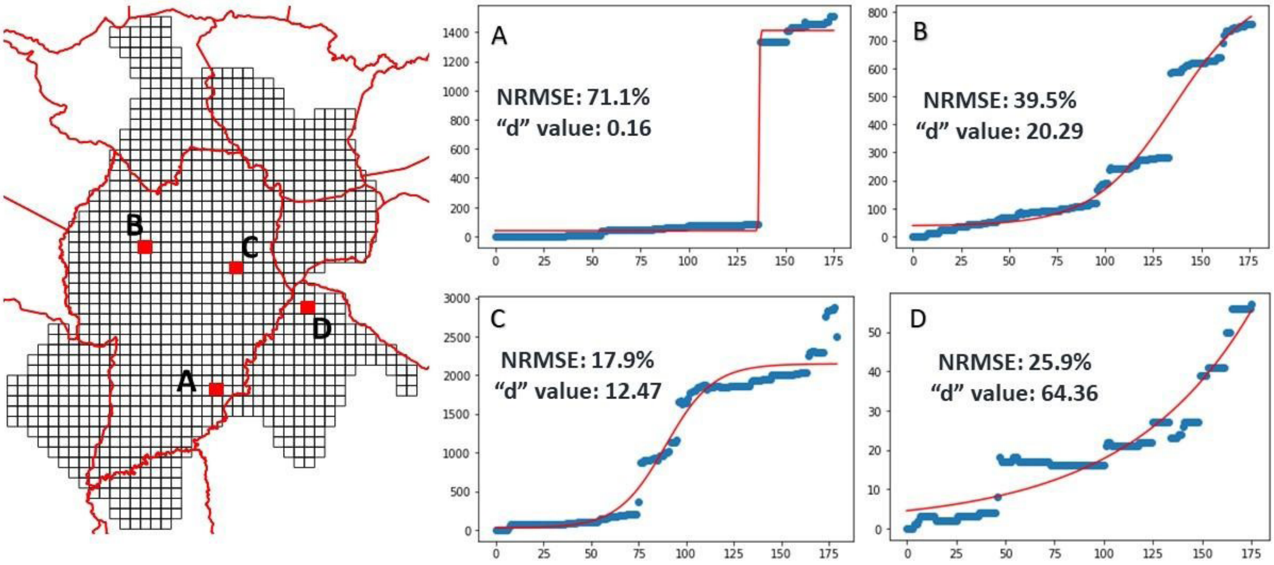


Figure 8: Behaviour of Logistic Regression in different cells evaluated.

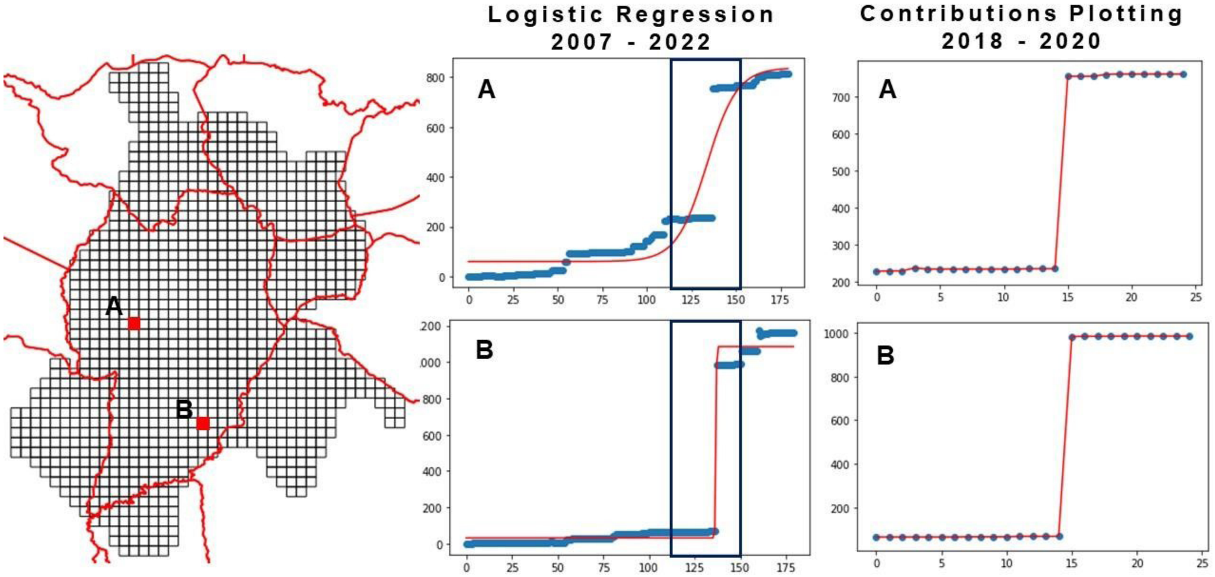


Figure 9: Behaviour examples of the abrupt jumps in the contribution patterns in different Curitiba cells.

As shown in Figure 9, it is possible to note that address imports influenced the total accumulated contributions. In region A, 519 features were added in a single month during the period (2018-2020). This corresponds to 63.4% of the total accumulated number of features by 2022. In region B, the number of contributions in a single month was 916, corresponding to more than 78% of the available features. In addition, the logistic regression presented an overflow message, indicating an abrupt jump.

Figure 10 presents the behavior of the logistic curve in different cells before and after the filtering of the OSM tag key "addr:street".

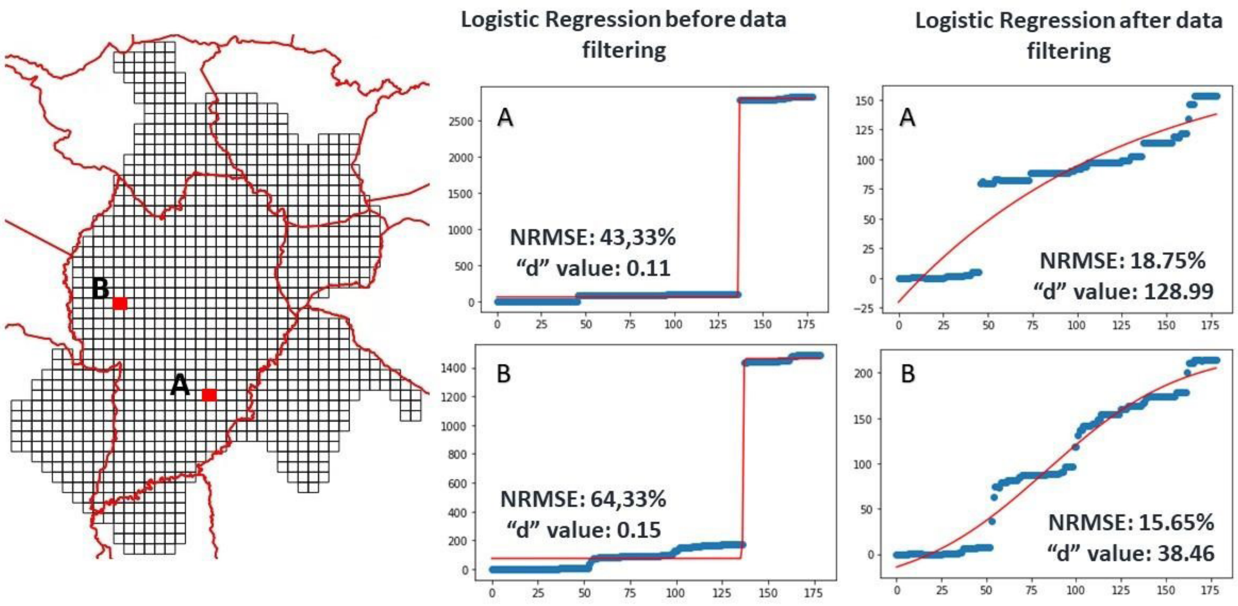


Figure 10: Logistic Regressions in different areas before and after filtering the address points.

From the graphs shown in Figure 10, it can be observed that from the data filtering, it was possible to obtain smoother curves in the regression calculation. An important aspect is that in graph A, even though a significant contribution represents a jump, it was not modelled in the regression after it occurred with gradual contributions. In addition, in the mathematical modelling that contemplates all the contributions, the high number of inserts in a single month made the regression indicate possible data saturation, both in graphs A and B. This aspect was not evident after filtering, indicating that the regions are still growing (Zone II). Another question was about the values obtained for parameter "d." After filtering the data, a significant increase in the value was observed. In this context, in steeper curves, the slope tends to zero (Figure 2), which allows the identification of areas of high growth in a short period or even measuring the behavior of the contributions. It was also possible to analyze regions influenced by significant contributions or events in addition to inserting address points. In this context, in Curitiba, cell clusters in which the contribution percentage in a single month is equivalent to more than 80% of the total amount.

In the region comprising Figure 11, it was observed that the described behaviour was buildings inserted simultaneously in a collective mapping activity in 2019, in which the set of changes was called *"Mapeo Colectivo realizado en el Marco del Master en Desarollo Urbano y Territorial (MDUT)Universidad Politecnica de Catalunya (UPC) #MapPyOSM."* This characteristic evidenced the abrupt growth observed in the logistic regression, in addition to having 1962 features inserted in a single month, equivalent to more than 90% of the accumulated total.

Therefore, when observing the distribution of contributions accumulated over time in heterogeneous regions, it is possible to observe patterns and behaviors that may help understand the dynamics of OSM contributions in a specific time and space. For example, the map in Figure 12 shows the spatial distribution of the month's percentage with the highest contributions concerning the accumulated total after filtering the OSM address tags.

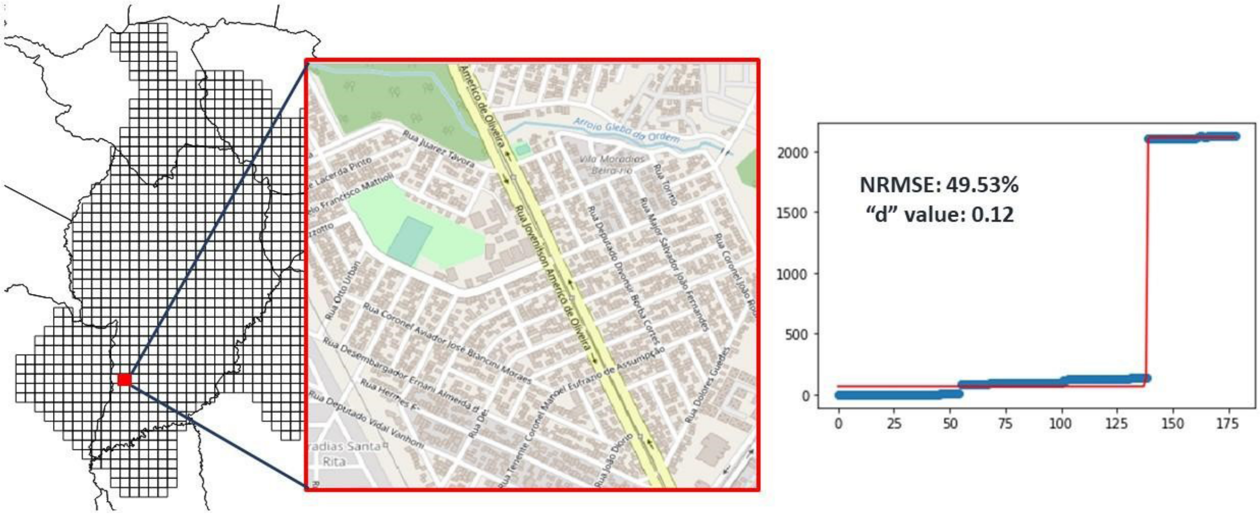


Figure 11: Cell in which the collective mapping activity took place.

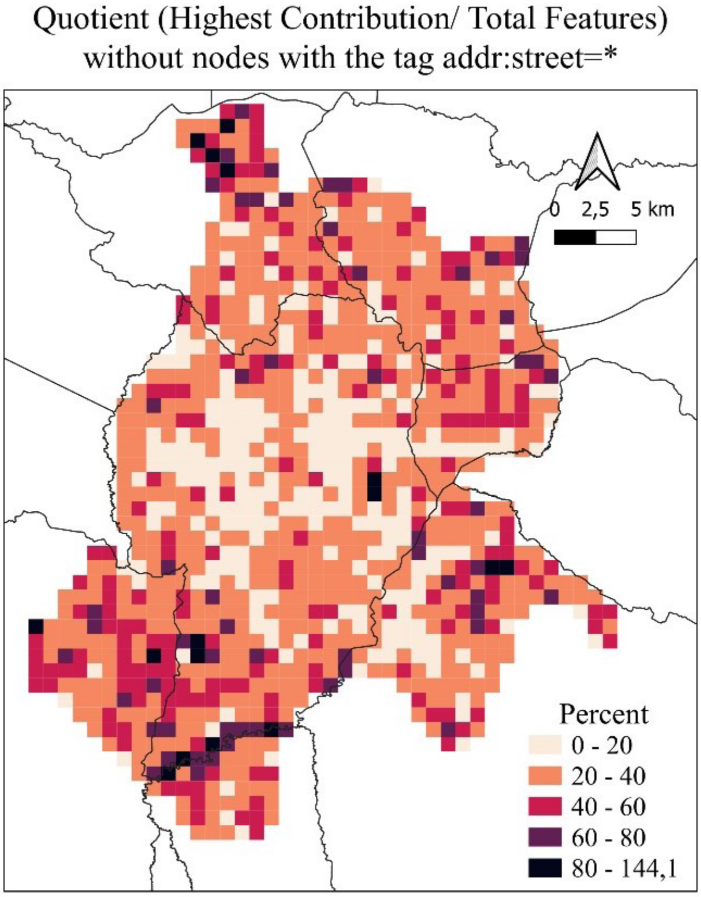

Quotient (Highest Contribution/ Total Features)
without nodes with the tag addr:street=*
0 2,5 5 km
Percent
0 - 20
20 - 40
40 - 60
60 - 80
80 - 144,1

Figure 12: Contributions in the highest month over the total (after filtering).

When comparing the answers obtained in Figure 12 to Figure 7B, it is possible to notice that there was a decrease in cells with peaks in the insertion of features in Curitiba, quantitatively demonstrating that the address points have a significant weight in the municipality analysis. This decrease made the results of Curitiba' cells equivalent to the total NUC. The predominant percentage was between 0% and 40%, which was obtained in more than 75% of the total cells.

A critical issue observed in Figure 13 was the presence of two cells in a specific region of Curitiba where the calculated percentage exceeded 100%. In these regions, there was a representative exclusion of features in the most recent month evaluated. Figure 13 shows the data patterns for this region. Graph A of Figure 13 presents the adjustment of the data from the logistic regression, and graph B corresponds to the number of features inserted or excluded by month. In the most recent month, 1319 features were excluded, and this value converges to the most significant number added (1265 features). This fact is most apparent in graph B, considering that there are two extreme peaks of addition and exclusion. However, the curve in graph A indicates that this is a typical growing cell (Zone II). The modelling of the curve has not yet adapted to this abrupt decrease, probably because the exclusion was in the later epochs of the study period and because of the short relative time gap between the contribution spike and its exclusion. Nevertheless, this characteristic corroborates the initial premise of data heterogeneity and that Spatio-temporal patterns can vary according to the evaluated period and region. All analysis results, with graphs for the 1074 cells of the region, can be interactively consulted at https://bit.ly/NUC_notebook .

The maps shown in Figures 14A and 14B depict the spatialization of the NRMS obtained in each cell before and after filtering the "addr:street" tag key.

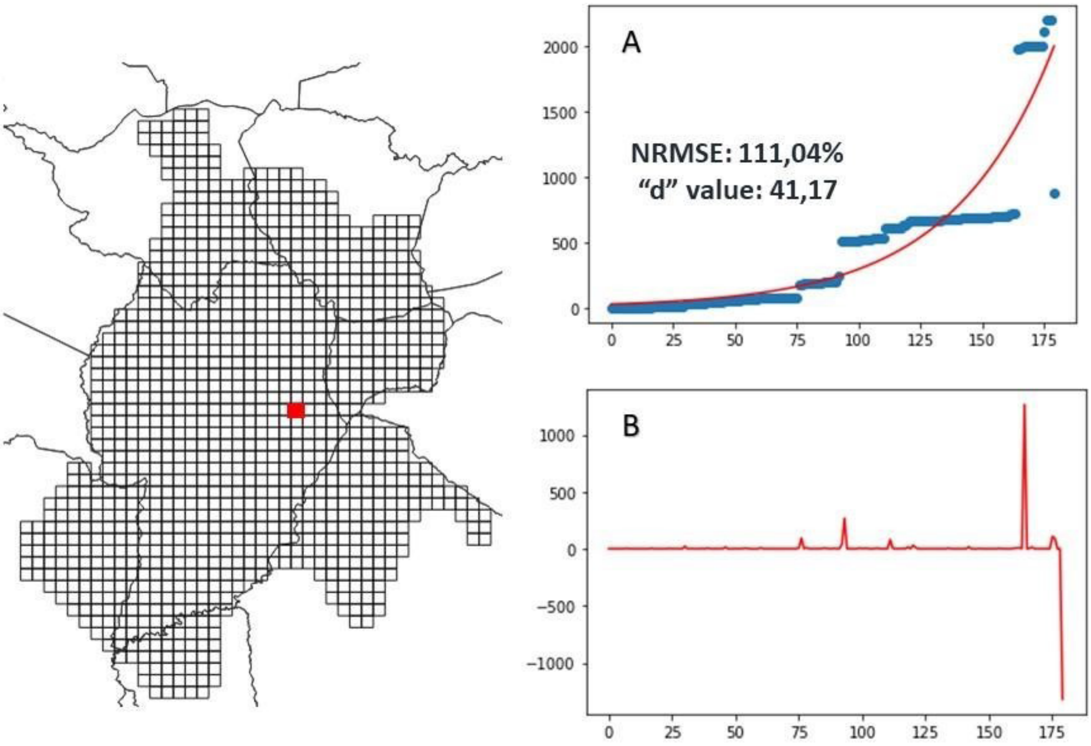


Figure 13: Cell in which representative exclusion of features was identified.

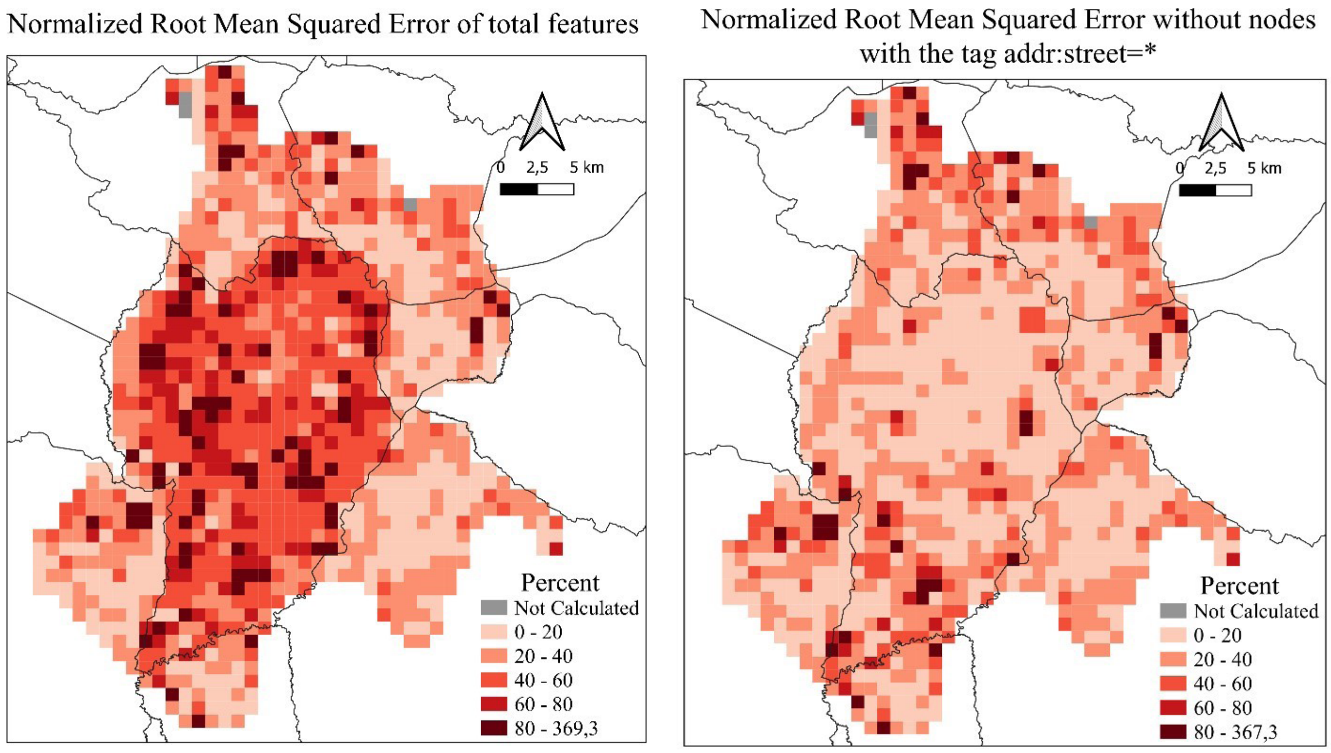


Figure 14A Figure 14B

Figure 14: NRMS obtained before (Figure 14A) and after filtering the addresses (Figure 14B).

When evaluating the regression quality through each cell NRMS, the first finding was that the address imports impacted the results obtained in Curitiba. In the analysis of all inserted features (Figure 14A), the mean and median NRMS were 41.247% and 35.7%, respectively. After filtering the addresses (Figure 14B), these values decreased to 27.536 and 20.6%, respectively. Therefore, there was an improvement in the regressions obtained, mainly due to the mass data importation to Curitiba. In Curitiba, we observed that NRMS was reduced after address filtering, and there were no significant variations in the other NUC regions. The mean and median NRMS values of the total number of features found in the cells intersecting Curitiba were 53.325% and 49.2%, respectively, while those in the other NUC regions were 29.546 and 22.2%, respectively. After address filtering, the results reached 25.455% and 19.1% in Curitiba and 29.552% and 22.2% in other locations, respectively.

In addition, three cells (represented in gray in Figure 14) failed in the NRMS calculation, as the median was zero. The curve failed to model the contributions in these regions because of the low amount of data input and accumulated growth without a typical pattern. As explained in the previous analysis of Figure 7, even if it is not an abrupt jump, regions with few features or unusual contribution behavior over time can also present this behavior. After address filtering, the overflow no longer appeared in Curitiba, which is comparable to the pattern observed in the other NUC regions (Figures 8 and 10). This last analysis also reinforces the use limitation of the presented method for areas where the number of contributions and their behavior are compatible with the logistic curve.

4. Conclusions and Recommendations

This study sought to develop an open-source methodology and tools to obtain the Spatio-temporal patterns of OSM and its mathematical modelling using logistic regression. The results indicate spatial data heterogeneity because the evolution of the contributions varies according to the region's characteristics. Among the causes of this variation is the existence of official open data for imports in OSM, initiatives to encourage collaborative mapping, and the characteristics of the population and urbanization of each area. The abrupt jumps detected are not negative, as they show areas where data growth is growing faster, either by the availability of open data or by focused actions of mapping communities in specific areas. However, such contributions differ from those of organic growth in areas where such phenomena do not occur. Furthermore, by verifying the parameters of steepness of the logistic regression, it was possible to identify these aspects and the areas affected by significant accumulated contributions in a single period over time. In addition, it was possible to notice that the contributions' dynamics can vary in the same region, which corroborates the prerogative that Spatio-temporal aspects in VGI can be directly related to the quality of geospatial data, especially regarding completeness and temporality.

Regarding the contribution modelling, it was possible to notice that the temporal aspects can be described through mathematical functions and by identifying the dynamics related to the evaluated region. In this context, for future work, it is recommended to continue research on the different factors related to the insertion of features in VGI over time, including other forms of mathematical modelling and other variables. In addition, it is recommended that the parameters found when obtaining the regressions be explored. In addition, there is a recommendation to perform spatial statistical analyses to classify the contribution patterns, which are the spatial characteristics of the distribution of such patterns and their relationships with other variables of the description of the territory and collaborative mapping.

ACKNOWLEDGEMENT

The authors would like to thank the Graduate Program in Geodetic Sciences at the Federal University of Paraná – Brazil. This study was carried out with the support of the Coordination for the Improvement of Higher Education Personnel (CAPES) – Funding Code 001.

AUTHOR´S CONTRIBUTION

Elias Nasr Naim Elias developed the study, computational application, experiments, and writing. Fabricio Rosa Amorim contributed to computational development, writing, and visualization results. Marcio Augusto Reolon Schmidt and Silvana Philippi Camboim contributed with supervision, formalization of analyses, methodological structuring, writing, and visualization results.

REFERENCES

Arsanjani, J. J. et al. 2015. An exploration of future patterns of the contributions to OpenStreetMap and development of a Contribution Index. Transactions in GIS, 19(6), pp. 896-914.

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

Brückner, J., Schott, M., Zipf, A., and Lautenbach, S. 2021. Assessing shop completeness in OpenStreetMap for two federal states in Germany. AGILE GIScience Series, 2, p. 20.

Camboim, S. P.; Bravo, J. V. M.; and Sluter, C. R. 2015. An investigation into the completeness of, and updates to, the Open Street Map data in a heterogeneous area in Brazil. ISPRS International Journal of Geo-Information, 4(3), p.1366-1388.

Cerqueira, E. V. and Diniz, A. M. A. 2022. Identifying centers and subcenters in the metropolitan region of Belo Horizonte through google places of interest. Mercator. Fortaleza, 21, p. e21012.

COMEC – Coordenação da Região Metropolitana de Curitiba. 2022. A Região Metropolitana de Curitiba. Available at:< https://www.comec.pr.gov.br/Pagina/Regiao-Metropolitana-de-Curitiba> [Accessed 13 October 2022].

Diretoria do Serviço Geográfico (DSG), 2015. Especificação Técnica para Controle de Qualidade de Dados Geoespaciais Vetoriais (ET-CQDG). Brasil.

Elias, E. N. N.; Fernandes, V. O.; Alixandrini Junior, M. J. and Schmidt, M. A. R. 2020. The quality of OpenStreetMap in a large metropolis in northeast Brazil: Preliminary assessment of geospatial data for road axes. Bulletin of Geodetic Sciences, 26(3), p. e2020012.

Grinberger, A. Y. et al. 2021. An analysis of the spatial and temporal distribution of large-scale data production events in OpenStreetMap. Transactions in GIS, 25(2), pp. 622-641.

Gröching, S., Brunauer, R., and Rehrl, K. 2014. Digging into the history of VGI data-sets: Results from a worldwide study on OpenStreetMap mapping activity. Journal of Location Based Services, 8(3), pp. 198–210.

Haklay, M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning B: Planning and design, 37(4), pp. 682-703.

Haklay, M., Basiouka, S., Antoniou, V., and Ather, A. 2010. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. The Cartographic Journal, 47(4), pp. 315–322.

IBGE – Instituto Brasileiro de Geografia e Estatística. 2022. Censo Demográfico 2022 - IBGE. Available at:<https://censo2022.ibge.gov.br/sobre/geografia-censitaria/enderecamento.html> [Accessed 29 December 2022].

IPPUC – Instituto de Pesquisa e Planejamento Urbano de Curitiba: Dados Geográficos. Available at:< http://www.ippuc.org.br/geodownloads/geo.html> [Accessed 13 October 2022].

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

Kitchin, R. 2015. The opportunities, challenges and risks of big data for official statistics, Statistical Journal of the IAOS, 31(3), p.471–481.

Le Guilcher, A., Olteanu-Raimond, A.-M. and Balde, M. B. 2022. Analysis of Massive Imports of Open Data in OpenStreetMap Database: A Study Case for Prance, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., V-4-2022, p.99–106

Nasiri, A. et al. 2018. Improving the quality of citizen contributed geodata through their historical contributions: The case of the road network in OpenStreetMap. ISPRS International Journal of Geo-Information, 7(7), p.253.

Paiva, C. dos A. and Camboim, S. P. 2021. A Dinâmica de Colaborações OpenStreetMap e sua Relação com as Atividades de Uso e Ocupação do Solo: um Estudo Segundo Zoneamento de Curitiba. Revista Brasileira de Cartografia, 73(1), p.73-87.

Paiva, C. dos A. and Camboim, S. P. 2022. Inference of positional accuracy of collaborative data from intrinsic parameters. Transactions in GIS, 26(4), pp. 1898–1913.

RAIS - Annual Report of Social Information. 2021. Available at:< http://www.rais.gov.br/sitio/sobre.jsf> [Accessed 29 December 2022].

Robinson, A. C. et al. 2017. Geospatial big data and cartography: Research challenges and opportunities for making maps that matter. International Journal of Cartography, 3(sup1), p.32–60.

Sehra, S. S., Singh, J. and Rai, H. S. 2017. Assessing OpenStreetMap data using intrinsic quality indicators: an extension to the QGIS processing toolbox. Future Internet, 9(2), p.15.

Silva, L. S. L. and Camboim, S. P. 2020 Brazilian NSDI ten years later: current overview, new challenges and propositions for national topographic mapping. Bulletin of Geodetic Sciences. 26(4): e2020018.

Teimoory, N., Abbaspour R. A. and Chehreghan A. 2021 Reliability extracted from the history file as an intrinsic indicator for assessing the quality of OpenStreetMap. Earth Science Informatics, 14 (3), p.1413–1432.

UN-GGIM. 2022. United Nations Integrated Geospatial Information Framework (UN-IGIF) Available at:<https://ggim.un.org/IGIF/part1.cshtml> [Accessed 29 December 2022].

Witt, R., Loos, L., and Zipf, A. 2021. Analysing the Impact of Large Data Imports in OpenStreetMap. ISPRS International Journal of Geo-Information, 10(8), p.528.

Tam, S.-M. and Van Halderen, G. 2020. The five V's, seven virtues and ten rules of big data engagement for official statistics. Statistical Journal of the IAOS, 36(2), p.423–433.

Zhang, H. and Malczewski, J. 2018. Accuracy Evaluation of the Canadian OpenStreetMap Road Networks. Internacional Journal Geospatial and Environmental Research, 5, p.1-14.