An automatic solution for determining the Big Geo Data characteristics and classification for thematic mapping

Jose A. D. Cacciatore [a],*, Claudia Robbi Sluter [a]

[a] *Federal University of Rio Grande do Sul, Graduate Program on Remote Sensing - josecacciatore89@gmail.com, robbi.sluter@ufrgs.br*

* Corresponding author

Keywords: Big Geo Data, digital cartography, automatic thematic mapping, Fisher-Jenks Algorithm

Abstract:

Today, our society generates billions of data daily, driven by the popularity of the internet and mobile devices (Robinson et al., 2017). The combination of Internet popularity and mobile devices has increased digital transactions, interactions on social networks, and sensors connected to the Internet of Things (IoT), generating a continuous flow and large volumes of data. When that data has some spatial attribute, we call it Big Geo Data (Goodchild, 2013). In the realm of Big Geo Data, determining the characteristics of the data, whether qualitative or quantitative, is essential for automating thematic mapping design. Traditional quantitative data classification methods include equal intervals, quantiles, standard deviation, and optimal classification, such as the Fisher-Jenks Algorithm (Slocum et al., 2009). However, those methods were proposed in a different technological reality: a limited quantity of data and a low data change rate over time. Big Geo Data is a vast source of heterogeneous data often generated in real-time and at a high temporal rate of change (Chang & Grady, 2019). In this modern context, our research aims to automatically collect geodata from the Internet, determine their characteristics, and define and apply a suitable classification method to generate thematic maps of quantitative data.

This paper presents research steps related to determining the geodata characteristics, separating qualitative and quantitative data, and automatically establishing the number of thematic classes and their numerical limits. To automate a quantitative Big Geo Data thematic mapping design, we propose a method comprising four steps: 1) collecting data, 2) separating them into quantitative and qualitative data groups, 3) saving quantitative data in a CSV file, and 4) classifying them using the Fisher-Jenks Algorithm. Initially, we tested the algorithm with simulated data generated at a frequency of 1s (Table 1), and later, we used the influenza database provided by the World Health Organization's Influenza Laboratory Surveillance Information - Influenza Virus Detections Reported to Flunet ILSI-IVDR, which can be understood as Big Geo Data.

raw random data generated per second	The quantity of data	NC	**GADF**
Start de algorithm	0	0	0
After 60 seconds	60	13	0.97
After 120 seconds	120	10	0.91
After 180 seconds	180	8	0.88
After 240 seconds	240	7	0.85
After 300 seconds	300	5	0.81
After 600 seconds	600	4	0.77
After 3600 seconds	3600	5	0.81
After 43200 seconds	43200	6	0.83
After 86400 seconds	86400	5	0.81

Table 1. The proposed algorithm's performance in a constant data flow over 24 hours. NC is the number of classes, and GADF is the goodness of absolute deviation fit.

Abstracts of the International Cartographic Association, 10, 29, 2025
32nd International Cartographic Conference (ICC 2025), 17–22 August 2025, Vancouver, Canada.
https://doi.org/10.5194/ica-abs-10-29-2025

Subsequently, we automatically separated the input data into quantitative or qualitative. In the third step, the quantitative data is saved in a CSV (Comma-Separated Values) file, which is chosen because this data format is compatible with traditional manipulation tools such as Excel, LibreOffice, ArcGIS, and QGIS. The compatibility with different software ensures interoperability with different digital solutions. Finally, we classified the data using the Fisher-Jenks Algorithm, and the choice of this method is justified because it considers the natural distribution of the data (Slocum et al., 2009), making it particularly suitable for Big Geo Data.

Our systematic approach organizes the information efficiently and prepares the dataset for use in the cartographic design of thematic maps. In this context, from data collection to identification of the type of data and its classification, our algorithm also automatically calculates the ranges of the classes to be represented in the maps. With the simulated data, the algorithm worked well for results of up to 10 classes. Over 10 classes, it is not stable, which means that the algorithm presents different results when running at different moments for the same data set. We achieved better results for 4 or 5 classes. Table 1 shows some of the results obtained for simulated data.

The algorithm worked well with simulated random data. There was some instability at the beginning; however, the more extensive the quantity of data, the better the result of calculating the number of classes. In this example, the result was five classes after 24 hours of data collection, totalling 86400 data (Table 1). These results showed us that the algorithm is efficient for defining the number of classes and, consequently, the numerical limits of the classes for data collected every other minute. We also classified the data from the Flunet ILSI-IVDR by country, making it possible to identify the eight subtypes of influenza in them and the subtype that most predominates in the population of each country. Our research results show that this proposed algorithm can be applied to identify and control disease outbreaks and other areas of science. The algorithm was developed using the Python programming language (PYTHON SOFTWARE FOUNDATION, 2024), in an Anaconda (ANACONDA, INC., 2020) environment, with the help of Jupyter Notebook (PROJECT JUPYTER, 2024) for prototyping and data analysis.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001.

References

ANACONDA, Inc. *Anaconda Distribution* (Version 2020) [software]. 2020. Available at: https://www.anaconda.com. Accessed on: May 14, 2025.

Chang, J. F., & Grady, N. 2019. Big Data: Storage, Sharing, and Security. In Handbook of Research on the Evolution of IT and the Rise of E-Society (pp. 193-212). IGI Global. DOI: 10.4018/978-1-5225-7214-5.ch010

Goodchild, M. F. 2013. The quality of big (geo) data. Dialogues in Human Geography, 3(3), 280–284 doi:10.1177/2043820613513392.

PROJECT JUPYTER. *Jupyter Notebook* [software]. 2024. Available at: https://jupyter.org. Accessed on: May 14, 2025.

PYTHON SOFTWARE FOUNDATION. *Python: Version 3.13.3* [software]. 2024. Available at: https://www.python.org. Accessed on: May 14, 2025.

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), pp. 32–60. doi: 10.1080/23729333.2016.1278151.

Slocum, T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. 2009. Thematic Cartography and Geovisualization. Pearson

https://doi.org/10.5194/ica-abs-10-29-2025