International Cartographic Association
Association Cartographique Internationale

Assessing the Suitability of Data Classification Methods for Choropleth Maps Depicting Population Distribution in South Africa

Lourens Snyman*, Serena Coetzee and Victoria Rautenbach

Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa,
lourens.snyman@up.ac.za, serena.coetzee@up.ac.za, victoria.rautenbach@gmail.com

* Corresponding author

Keywords: data classification methods, choropleth map, population distribution, South Africa

Abstract:

Techniques for data visualization, or rather geospatial data visualization evolved rapidly over the past decade or so. Built-in spatial analysis and visualizations, such as choropleth maps, kernel density estimation (KDE) heat maps, firefly maps, dot maps, graduated symbols, point-density or isochrones to mention a few, are standard tools nowadays in most open source and proprietary geographic information system (GIS) applications, enabling users to quickly visualize spatial patterns in data. With the increased processing capability of desktop computers, high-speed internet connections for cloud computing, analysing and displaying large volumes of geographic data became easier and much faster.

Current proprietary and open source GIS software development companies or communities are continuously simplifying their applications allowing non-GIS professionals to execute advanced spatial queries and visualization techniques with the click of a few buttons. The user interfaces (UI) are developed in such a way that the end user (people with or without GIS knowledge) is guided through logical steps to meet their objective. Whether it is to create a topographic locality map or choropleth map depicting population distribution, a step-wise approach is predefined to guide the user through the analysis and visualisation processes. Although these steps enable non-GIS users to create informative maps, the software cannot verify that the most appropriate method was used to convey the message or tell the story. For this, humans are needed.

Choropleth maps are one of the oldest, but still one of the most popular (or frequently used) techniques to visualise quantitative data in a GIS (Tyner, 2014). Slocum et al. noted that a choropleth map is “the most commonly used (and abused) method of thematic mapping” (Slocum et al., 2014). A choropleth map group (or combines) observations into classes based on a data classification method. Again, most GIS software applications include a range of data classification methods for choropleth maps, making it easy for users to select and apply a method from a drop-down menu. That being said, choosing a suitable, or most effective data classification method for a specific dataset remains a challenge.

South Africa is characterised by an uneven geographic distribution of people. Besides densely populated central business areas, most municipalities also have scattered pockets of informal or semi-informal settlements situated in peri-urban areas or just outside the city centre. Ideally, an effective data classification method should not only highlight (or accentuate) primary high-density locations such as city centres and surroundings, but also populated secondary and tertiary locations like townships and informal settlements.

The aim of this research was to assess the suitability of data classification methods for choropleth maps to effectively visualize population distribution in South Africa. Through an online questionnaire, participants were asked to answer a series of questions related to supply (service centres) and demand (population distribution) in four different municipalities across South Africa, which would inform decisions about optimizing service delivery. The questionnaire was designed in Qualtrics (https://www.qualtrics.com/uk/) and consisted of 48 questions. For each question, participants were required to identify one or more locations on a map. The XY pixel value of responses was saved in the backend, allowing validation of correct and incorrect responses.

Although various data classification methods are described in literature, only those available in either ArcGIS Pro or QGIS (since they are the most popular and frequently used GIS software in South Africa) were considered for the study. With the exception of manual and defined interval (where the user is required to set custom/manual limits for

each class), seven data classification methods were initially identified namely equal interval, geometric interval, logarithmic scale, natural breaks (jenks), pretty breaks, quantiles and standard deviation.

Prior to the design of the online questionnaire, data classification methods were evaluated based on a visual inspection, as well as recommendations from the literature. Factors such as the skewness of the data distribution and accuracy index calculations were considered (Jenks and Caspall 1971, Robinson 1984). From the seven data classification methods, only geometric interval, logarithmic scale, natural breaks (jenks) and quantiles were eventually selected for the user study.

Participation of students from the University of Pretoria was completely voluntary. Students were invited to participate, as they are representatives of the future workforce. A few lucky draw cash prizes were given as an incentive to participants who completed the survey. From the 107 participants, most were registered for the B.Ed programme (Education) at the time of the survey (40%), followed by BSc Geography and Environmental Science and BSc Geoinformatics with 14% and 11% respectively. The age of participants ranged from 18 to 27, with the majority in the age group 19 to 21. Most participants were female (64%), followed by males (35%), and one student preferred not to indicate a specific gender. The majority of participants (90.6%) took geography as a subject at school. From those with geography at school, 43.2% were enrolled for the B.Ed. academic programme. Most participants are 2nd year students (78.5%) followed by 3rd year students (12.1%).

Figure 1 shows an example of responses (red dots). Participants were asked to identify areas with high population density (demand), where service centres (supply) are needed. The choropleth map for the City of Tshwane Metropolitan Municipality shows population density based on the quantiles data classification method grouped into five class intervals.

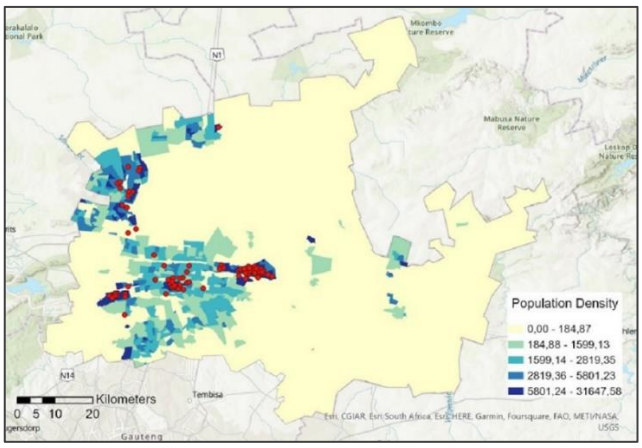

Figure 1. Choropleth map showing population density for the City of Tshwane Metropolitan Municipality

Initial tests of the survey results include the Kruskal-Wallis test to determine if there are statistically significant differences between the accuracy scores (percentage correct answers) of participants based on the four data classification methods. A significance score of <0.001 confirms that at least one of the data classification methods has a different distribution compared to the others.

Overall, participants did well with an average percentage accuracy of 89.9%. Four participants scored 100% and the minimum recorder score was 64.6%. The quantiles data classification method scored the highest percentage accuracy of 92.3%, followed by natural breaks (jenks) and geometric interval with 91.2% and 88.8% respectively.

A more in-depth explanation of the methodology and results will be presented at the conference which will also comprise a list of significant predictors influencing correct and incorrect answers. Future research could include the evaluation of additional data classification methods with varying numbers of class intervals.

References

Jenks, G. F. and F. C. Caspall (1971). "Error on choroplethic maps: definition, measurement, reduction." Annals of the Association of American Geographers **61**(2): 217-244.

Robinson, A. H. (1984). Elements of cartography. New York, Wiley.

Slocum, T. A., et al. (2014). Thematic Cartography and Geovisualization, Pearson.

Tyner, J. A. (2014). Principles of map design, Guilford Publications.

https://doi.org/10.5194/ica-abs-7-160-2024