Rule-Based Geospatial Visualisation Recommendation

Anthony Joseph Jungo, Kaiza Kunonu Ilomo, John Waithaka
Carnegie Mellon University
Kigali, Rwanda
{ajosephj, kilomo, jwaithak}@andrew.cmu.edu

I. Introduction

We intend to investigate the automation of effective geospatial data visualisation using rule-based systems. An effective data visualisation is one that accurately represents useful patterns in data and efficiently communicates these patterns [1]. In this study, we focus on the accuracy of geospatial visualisation, specifically looking at automatically detecting accuracy-reducing overcrowding in point maps. Overcrowding, or overplotting, is a common problem in data visualization. It occurs when too much data is displayed in a small area, causing visual noise and obscured information rather than accurate information conveyance [2], [3]. Rule-based systems, as opposed to machine learning-based systems, are systems that encode domain knowledge into a program to automate or assist in domain expert activities. In our context, the domain knowledge is geospatial data visualisation.

A. Problem Statement

Data visualisation is a fundamental part of data analysis. It enables the communication of patterns in large volumes of data, therefore enabling data-driven decision-making. However, effectively communicating patterns in data through visualizations is not easy. There are many design choices involved, each of which has consequences on the effectiveness of the resulting visualisation. Choosing an optimal combination of design choices for visualising a given pattern, therefore, often requires considerable skill. This makes data visualisation and consequently data analysis, inaccessible in the absence of expertise. However, even in the presence of expertise, effective data visualisation is costly and time-consuming, since it is often a manual trial-and-error process [4], [5].

B. Research Question

We seek to answer the question, how can a rule-based system be developed to automatically create effective geospatial visualisations by making optimal design choices based on the characteristics of the data being visualised? We specifically look at automating the detection of overcrowding in point maps to inform the design choice of whether a point map would be effective for a given dataset.

C. Intended Audience

This paper is intended for people interested in the development of automated geospatial data visualization systems. Further, since the proposed methods may apply to other types of data visualization, it is also relevant for those interested in the development of automated systems for general data visualization.

The intended users of the proposed systems are people with influence over managerial or executive decisions but no access to data visualisation skills. Such a system would help them gain and communication insights from geospatial data to make informed decisions, even in the absence of data analysis experts. We assume that such people are sufficiently competent to interpret geospatial visualisations and have enough computer skills to operate the proposed system and handle computer files.

D. Scope

This research is focused on automating the decision-making process of determining whether point maps are an effective visualisation type for a particular dataset. While many design choices impact the effectiveness of data visualisations, we concentrate specifically on this aspect.

II. Importance and Prior Work

The benefits of data analysis for decision-making in organisations [6], healthcare [7] and public administration are increasingly well known. However, these benefits are highly dependent on data visualisation, which is a fundamental part of data analysis [8]. However effective data visualisation is a hard task and often requires expertise [9], which is costly. This makes it and, consequently, data analysis inaccessible to many.

In response to this, researchers have been attempting to develop visualisation recommendation systems since as early as the 1980s [10]. Visualisation recommendation systems automatically suggest useful and communicative visualisations for a given dataset, to make effective data visualisation easier and more accessible [11]. These works are based on the findings that the characteristics of a dataset affect how it can and should be visualised.

A. Prior Work

Prior work on visualisation recommendation systems can be categorised at a high level into work on rule-based systems and the more recent Machine learning-based systems. Machine learning-based visualisation recommendation systems (e.g., [9], [12], [13]) involve learning a function that maps datasets to effective visualisations using a large corpus of datasets and their corresponding visualisations. Some of these works have

resulted in systems with good performance, however, a significant amount of effort and cost is required to acquire and prepare sufficient training and testing data for a generalisable ML-based recommendation system [9]. Further, commonly used machine learning models, such as neural networks [9], [13], are criticised for being hard to interpret. That is, understanding the reasoning behind their recommendations can be challenging [11].

Rule-based visualisation recommendation systems (e.g, [10], [14], [15]) implement principles of data visualisation derived from literature on effective visualisation. Mackinlay's foundational work APT [10], for example, implements principles drawn from Cleveland and McGill's Graphical Perception [16] on the perceptual accuracy of interpretation of visual encodings of quantitative data. These types of systems require minimal data since they do not learn from data, but instead are pre-programmed with the necessary knowledge. Consequently, they do not incur the data acquisition costs associated with ML-based systems. Further, since their output is derived from principles implemented by the system developers, these systems' outputs are highly interpretable. However, the challenge with these systems lies in implementing these principles and measuring the factors affecting these principles. Our research takes the approach of rule-based systems.

B. *Effective Data Visualisation*

To recommend effective visualisations, rule-based visualisation recommendation systems implement principles of effective visualisation to identify good visualisations from the possible options for a given dataset. The early foundational work on effective visualisation by Bertin [17] and Cleveland and McGill [16] measure visualisation effectiveness by ranking the effectiveness of individual visual encodings (e.g., x-position, size) at encoding a certain variable type (e.g., categorical or quantitative variables). Later work improved on this by looking at other characteristics of a dataset besides variable types. These include things such as cardinality and distribution of variables [18], [19]. Other works focus on specific elements of a visualisation such as overcrowding and its degrading effect on the effectiveness of a visualisation [2].

Overcrowding in data visualisation is a common problem today given the increase in the size of datasets [3]. Bertini and Santucci [2] propose a number of metrics for measuring overcrowding in scatter plots. These include the crowded points to total points ratio (henceforth crowded points ratio), where crowded points refer to instances where multiple points collide within a very small portion of the display area. The threshold for the number of colliding points that constitute a crowd (k) is an adjustable parameter. Further, the threshold for the acceptable crowded points to total points ratio is also an adjustable parameter. Our research seeks to apply this work on scatter plots to point map visualisation

III. Methodology

We divide this work into the following sub-problems.

- What are the principles of creating effective geospatial visualisations?
- How can a system that automatically enforces geospatial visualisation principles, specifically regarding overcrowding in point maps, be developed?
- How can the developed system be tested?

A. *What are the principles of creating effective geospatial visualisations?*

Our initial step involved identifying principles and guidelines for effective geospatial visualisation. We accomplished this by searching and reviewing relevant literature. Besides reviewing the literature, we also experimented with different identified visual encodings on geospatial datasets to identify obvious principles. These experiments were done using Python on a Jupiter Notebook, and the geospatial datasets were sourced from The Africa GeoPortal [20] open geospatial data portal.

B. *How can a system that automatically enforces geospatial visualisation principles, specifically regarding overcrowding in point maps, be developed?*

To enforce visualisation principles on overcrowding we first developed a program to measure anticipated overcrowding. This program used the crowded points ratio metric proposed by Bertini and Santucci [2]. It takes a geospatial dataset and dimensions of the map to be plotted as inputs. It forms a virtual map of the given dimensions. It divides this map into grid cells, where each cell has the same area as the points in the point map to be plotted. It then places each data sample in the geospatial dataset onto the virtual map in its correct spatial position. It then counts the number of data samples within each cell. With these counts, it can query the cells that contain a data sample count that exceeds the overcrowding threshold, k. And with these overcrowded cells, it can get the number of data samples in a crowd by summing up the counts of the overcrowded cells. This program was developed using Python due to its large ecosystem of libraries for data manipulation and creating visualisations.

C. *How can the developed system be tested?*

To test the system, open geospatial data was sourced from The Africa GeoPortal [20] and Geodatasets [21]. Ten datasets were retrieved and visualised using point maps. By visual observation, we determined that six of these were degraded due to overcrowding while the remaining four had no or little degradation due to overcrowding. We ran the program on these six of these datasets (four with too much overcrowding and two without), manually tuning k and the acceptable crowded points ratio. After this, we ran the tuned program on the remaining four datasets to test whether it would correctly identify too much overcrowding and recommend or discourage the use of point maps.

IV. Results and Discussion

A. *Principles of Effective Geospatial Visualisation*

The principles and guidelines for effective data visualisation recommend visualisation design decisions for a given dataset and user task. These recommendations are meant

to maximise the usefulness and communicativeness of the visualisations created. The general design decisions involved in the design of a geospatial visualisation are map type (e.g., point maps, heatmaps and choropleths), visual encodings (e.g., colour, symbols, labels, area) and map distortion (cartograms). The effectiveness of these decisions depends on the characteristics of the dataset being visualised and the user's goal or task. Design decisions are highly interdependent - one decision constrains decisions in other dimensions of the design.

One category of design decisions involves the type of map used. A popular map type is choropleths. Choropleths present aggregated data about a geographic area using colour or pattern visual encodings [22]. These maps are good at visualising categorical data about geographic regions, for example, the dominant spoken language per country, using colour or pattern fills to encode the categorical variable. Choropleths are also good at visualising quantitative variables using colour gradient or shade. A disadvantage of choropleths is that sector areas are often interpreted as representing a quantitative variable, thus misleading users [23]. Previous works propose cartograms and ensemble coding to solve this problem [23], [24]. This research, however, is not concerned with choropleths or cartograms.

Hexagonal density maps, like choropleths, visualise aggregated data. However, the area over which they aggregate data is not geographic sectors but sufficiently small hexagons. A map is tiled with non-overlapping hexagons, leaving no gaps [25]. The point data that fall into a tile are aggregated, for example by count or voting, and the aggregated data is visually encoded using colour. They are effective at visualising both quantitative and categorical variables. This map type is often used as an alternative to point maps affected by overcrowding.

Heatmaps visualise the density of phenomena based on point data [26]. Areas with a denser cluster of points are emphasised. Density information is visually encoded using colour hue or intensity. Heatmaps do not visualise any data variable, only the density of points.

Point maps are a basic geospatial visualisation type that show precise locations of entities like health centres in a country. They are effective at visualising both categorical and quantitative variables as well as point density. However, they are vulnerable to overcrowding, which obscures information and reduces their effectiveness [3]. Several techniques for reducing overcrowding have been proposed, including clustering, sampling, filtering and using alternative visualisation types [3]. We focus on the latter technique, specifically with choropleths, hexagon maps and heatmaps. Although these types of visualizations effectively address overcrowding, each gives away certain information. For example, choropleths may aggregate data within large spatial areas and therefore lose finer details present at more granular levels, and using heatmaps makes it impossible to visualise multivariate data. Therefore, the design decision of which alternate visualisation to use depends on the information a user intends to read from the visualisation, i.e., the user's task.

Amar et al. [27] identified ten low-level user tasks. We will consider only four of these, which are filtering (i.e., identifying data samples that satisfy a given condition), finding extremums, identifying outliers and identifying pattern. Each of these is effectively visualised by colour encoding [28]. For example, colour in heatmaps shows the locations with maximum and minimum density. Hexagons and heatmaps are effective for all four of these tasks, however, heatmaps are limited to visualising density data. Choropleths, due to their large area of aggregation are not as effective as hexagons and heatmaps for identifying outliers and pattern recognition.

As a proof of concept, we use the mentioned user tasks and visualisation types to develop the following decision system (Fig. 1) for deciding how to solve the problem of an overcrowded point map.

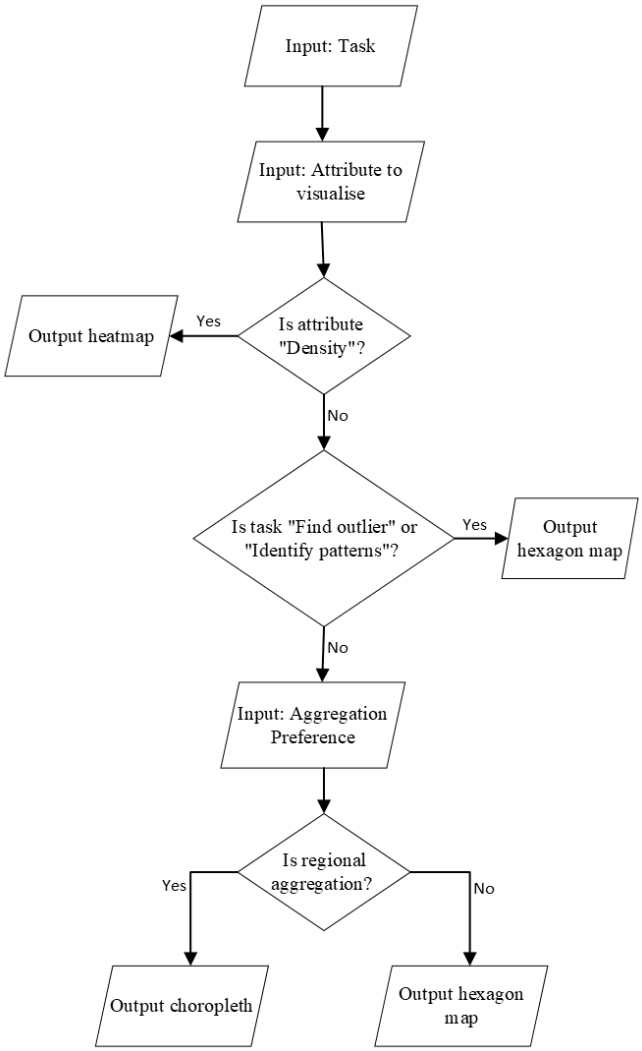

Fig. 1. Flowchart on choosing alternative visualisation to an overcrowded map

However, this depends on the identification of overcrowding.

B. Automatically Identifying Overcrowding in Point Maps

…

REFERENCES

[1] Y. Zhu, ‘Measuring Effective Data Visualization’, in Advances in Visual Computing, G. Bebis, R. Boyle, B. Parvin, D. Koracin, N. Paragios, S.-M. Tanveer, T. Ju, Z. Liu, S. Coquillart, C. Cruz-Neira, T. Müller, and T. Malzbender, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, pp. 652–661. doi: 10.1007/978-3-540-76856-2_64.

[2] E. Bertini and G. Santucci, ‘Give Chance a Chance: Modeling Density to Enhance Scatter Plot Quality through Random Data Sampling’, Inf. Vis., vol. 5, no. 2, pp. 95–110, Jun. 2006, doi: 10.1057/palgrave.ivs.9500122.

[3] G. Ellis and A. Dix, ‘A Taxonomy of Clutter Reduction for Information Visualisation’, IEEE Trans. Vis. Comput. Graph., vol. 13, no. 6, pp. 1216–1223, Nov. 2007, doi: 10.1109/TVCG.2007.70535.

[4] X. Qin, Y. Luo, N. Tang, and G. Li, ‘Making data visualization more efficient and effective: a survey’, VLDB J., vol. 29, no. 1, pp. 93–117, Jan. 2020, doi: 10.1007/s00778-019-00588-3.

[5] T. Siddiqui, A. Kim, J. Lee, K. Karahalios, and A. Parameswaran, ‘Effortless Data Exploration with zenvisage: An Expressive and Interactive Visual Analytics System’. arXiv, Jan. 04, 2018. Accessed: Mar. 28, 2024. [Online]. Available: http://arxiv.org/abs/1604.03583

[6] E. Brynjolfsson, L. M. Hitt, and H. H. Kim, ‘Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?’ Rochester, NY, Apr. 22, 2011. doi: 10.2139/ssrn.1819486.

[7] W. Raghupathi and V. Raghupathi, ‘Big data analytics in healthcare: promise and potential’, Health Inf. Sci. Syst., vol. 2, no. 1, p. 3, Feb. 2014, doi: 10.1186/2047-2501-2-3.

[8] W. S. Cleveland, Visualizing Data, 1st edition. Murray Hill, N.J. : Summit, N.J: Hobart Pr, 1993.

[9] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo, ‘VizML: A Machine Learning Approach to Visualization Recommendation’, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, in CHI ’19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12. doi: 10.1145/3290605.3300358.

[10] J. Mackinlay, ‘Automating the design of graphical presentations of relational information’, ACM Trans. Graph., vol. 5, no. 2, pp. 110–141, Apr. 1986, doi: 10.1145/22949.22950.

[11] H. Li, Y. Wang, S. Zhang, Y. Song, and H. Qu, ‘KG4Vis: A Knowledge Graph-Based Approach for Visualization Recommendation’, IEEE Trans. Vis. Comput. Graph., vol. 28, no. 1, pp. 195–205, Jan. 2022, doi: 10.1109/TVCG.2021.3114863.

[12] Y. Luo, X. Qin, N. Tang, and G. Li, ‘DeepEye: Towards Automatic Data Visualization’, in 2018 IEEE 34th International Conference on Data Engineering (ICDE), Apr. 2018, pp. 101–112. doi: 10.1109/ICDE.2018.00019.

[13] V. Dibia and Ç. Demiralp, ‘Data2Vis: Automatic Generation of Data Visualizations Using Sequence-to-Sequence Recurrent Neural Networks’, IEEE Comput. Graph. Appl., vol. 39, no. 5, pp. 33–46, Sep. 2019, doi: 10.1109/MCG.2019.2924636.

[14] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, ‘Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations’, IEEE Trans. Vis. Comput. Graph., vol. 22, no. 1, pp. 649–658, Jan. 2016, doi: 10.1109/TVCG.2015.2467191.

[15] D. Moritz et al., ‘Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco’, IEEE Trans. Vis. Comput. Graph., vol. 25, no. 1, pp. 438–448, 2019, doi: 10.1109/TVCG.2018.2865240.

[16] W. S. Cleveland and R. McGill, ‘Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods’, J. Am. Stat. Assoc., vol. 79, no. 387, pp. 531–554, 1984, doi: 10.2307/2288400.

[17] J. Bertin, Semiology of Graphics: Diagrams, Networks, Maps. ESRI Press, 2011.

[18] Y. Kim and J. Heer, ‘Assessing Effects of Task and Data Distribution on the Effectiveness of Visual Encodings’, Comput. Graph. Forum, vol. 37, no. 3, pp. 157–167, 2018, doi: 10.1111/cgf.13409.

[19] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri, ‘Four types of ensemble coding in data visualizations’, J. Vis., vol. 16, no. 5, p. 11, Mar. 2016, doi: 10.1167/16.5.11.

[20] ‘The Africa GeoPortal’. Accessed: Apr. 25, 2024. [Online]. Available: https://www.africageoportal.com/

[21] ‘Geodatasets’. Accessed: May 05, 2024. [Online]. Available: https://geodatasets.readthedocs.io/en/latest/

[22] M. Shaito and R. Elmasri, ‘Map Visualization using Spatial and Spatio-Temporal Data: Application to COVID-19 Data’, in Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference, in PETRA ’21. New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 284–291. doi: 10.1145/3453892.3461336.

[23] L. Besancon, M. Cooper, and A. Ynnerman, ‘An Evaluation of Visualization Methods for Population Statistics Based on Choropleth Maps’.

[24] D. Dorling, ‘Area Cartograms : Their Use and Creation’, Concepts Tech. Mod. Geogr. CATMOG, 1996, Accessed: May 05, 2024. [Online]. Available: https://cir.nii.ac.jp/crid/1574231875022545408

[25] T. Wang, ‘Adaptive Tessellation Mapping (ATM) for Spatial Data Mining’, Int. J. Mach. Learn. Comput., vol. 4, pp. 478–482, Jan. 2015, doi: 10.7763/IJMLC.2014.V6.458.

[26] W. Pokojski, T. Panecki, and K. Słomska-Przech, ‘Cartographic visualization of density: exploring the opportunities and constraints of Heat Maps’, Pol. Cartogr. Rev., vol. 53, no. 1, pp. 21–36, Jan. 2021, doi: 10.2478/pcr-2021-0003.

[27] R. Amar, J. Eagan, and J. Stasko, ‘Low-level components of analytic activity in information visualization’, in IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., Oct. 2005, pp. 111–117. doi: 10.1109/INFVIS.2005.1532136.

[28] G. J. Quadri and P. Rosen, ‘A Survey of Perception-Based Visualization Studies by Task’, IEEE Trans. Vis. Comput. Graph., vol. 28, no. 12, pp. 5026–5048, Dec. 2022, doi: 10.1109/TVCG.2021.3098240.