# 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 visualizations, we concentrate specifically on this aspect.

## II. Importance and Prior Work

Previous works in automated visualization recommendation systems fall into two categories: machine learning-based systems and rule-based systems.

Machine learning (ML)-based systems [6], [7], [8] involve learning a function that maps data sets to effective visualisations from a large corpus of labelled data. Hu et al. [6], for example, used a neural network trained on a corpus of about 2 million samples. The resulting model performed comparably to people with some experience and knowledge in visualising data. Despite the good performance of these systems, a significant amount of effort and cost is required to acquire and prepare sufficient training and testing data for a generalisable recommendation model [6]. In addition, commonly used machine learning models, such as neural networks [6], [8], often produce recommendations that are difficult to interpret. That is, understanding the reasoning behind their output can be challenging [9].

Rule-based visualisation systems follow from Mackinlay's pioneering A Presentation Tool (APT) [10]. These include Voyager [11], Draco [12], Show Me [13] and Sage [14]. These systems codify data visualisation principles as rules in a program. These rules define how data, having certain characteristics, can or should be visually encoded. One class of rules defines the valid and invalid data variable-to-encoding

mappings; Hu et al. [6] give an example of an invalid mapping as encoding a categorical variable with the y-position of a line chart. We are not concerned with this class of rules in this research. The other class of rules define the encodings that are most effective for data with certain characteristics. These rules are intended to optimise certain properties of visualisations, such as perceptual effectiveness [10], [11].

Notably, all prior works identified, consider only typical tabular data and visualizations like bar charts, line charts and scatter plots, and not geospatial data and visualizations. This raises the question of whether such works apply to geospatial data and visualization.

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, 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 is an adjustable parameter. Further, the threshold for the acceptable crowded points to total points ratio is also an adjustable parameter.

## 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?

## 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] 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*, (CHI '19), New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12. doi: 10.1145/3290605.3300358.

[7] 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.

[8] 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.

[9] 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.

[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] 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.

[12] 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.

[13] J. Mackinlay, P. Hanrahan, and C. Stolte, 'Show Me: Automatic Presentation for Visual Analysis,' *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1137–1144, Nov. 2007, doi: 10.1109/TVCG.2007.70594.

[14] S. F. Roth, J. Kolojejchick, J. Mattis, and J. Goldstein, 'Interactive graphic design using automatic presentation knowledge,' in *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence* (CHI '94), Boston, Massachusetts, United States: ACM Press, 1994, pp. 112–117. doi: 10.1145/191666.191719.