Rule-Based Geospatial Visualization Recommendation

Anthony Joseph Jungo, Kaiza Kunonu Ilomo, John Waithaka

Carnegie Mellon University

Kigali, Rwanda

{ajosephj,kilomo,jwaithak}@andrew.cmu.edu

# Introduction

We intend to investigate the effectiveness of rule-based systems in recommending effective geospatial data visualizations. An effective data visualisation is one that accurately represents useful patterns in data and efficiently communicatesthese patterns [1]. Following Hu et al. [2], we define the act of data visualization as the process of making design choices, based on the properties of a dataset, that maximize the effectiveness of the resulting visualization. Therefore, the recommendation of effective data visualisations is identifying a set of design choices that result in one or multiple effective visualizations for a given dataset. We focus on the visualization of geospatial data, which is data indexed with geographic coordinates [3]. Rule-based systems encode domain knowledge into a program to automate or assist in domain expert activities. In our context, the domain knowledge is geospatial data visualisation.

## Problem Statement

Data visualisation is a fundamental part of data analysis. It enables the identification and communication of patterns extracted from large volumes of data, therefore enabling data-driven decision-making. However, identifying and effectively communicating patterns in data through visualizations is not easy. There are innumerable combinations of design choices each good for revealing only certain patterns. Choosing the most effective visualization to reveal hidden patterns in a dataset is often not a simple task, even for experts. Data visualization and analysis are, therefore, largely inaccessible in the absence of data analysis expertise. However, even where there is expertise, the manual process is often time-consuming, laborious and costly since the expert will usually manually generate many visualizations in search of the most effective one.

## Research Question

We seek to answer the question, *how accurate are rule-based systems at finding useful and communicative visualisations for given geospatial datasets?* Accuracy is defined as the fraction of useful and communicative visualizations identified out of those identified by a knowledgeable human. It is also inversely proportional to the number of visualisations it deems useful and communicative, which a knowledgeable human deems not so.

## 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 analysis skills. Such a system could help them gain 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. Further, the proposed solution may be useful to data analytics experts by making the data exploration phase faster through automating the search for useful visualizations and reducing the need for manual trial and error search.

## Scope

This work will only consider enough geospatial data visualisation knowledge and design decisions to test the effectiveness of rule-based systems in recommending effective geospatial data visualizations. We will not attempt to develop a comprehensive system considering all known geospatial visualisation principles and design decisions. Further, we use small datasets to avoid the complexity of working with very large combinatorial spaces of design decisions and dataset subsets.

# Importance and Prior Work

## 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 [2], [4], [5] involve learning a function that maps data sets to effective visualisations from a large corpus of labelled data. Hu et al. [2], 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 [2]. In addition, commonly used machine learning models, such as neural networks [2], [5], often produce recommendations that are difficult to interpret. That is, understanding the reasoning behind their output can be  
challenging [6].

Rule-based visualisation recommendation systems follow from Mackinlay’s pioneering A Presentation Tool (APT) [7]. These include Voyager [8], Draco [9], Show Me [10] and Sage [11]. These systems implement data visualisation principles derived from work such as Bertin’s [12] and Cleveland and McGill’s [13] as rules. These rules define how data variables, having certain characteristics, can or should be visually  
encoded. **Visual encodings** are the representations of data variables in a visualisation. In a scatter plot of cars, for example, the y-position of a point may encode the price, the x-position, fuel consumption, and the shape of the point, the car model. One class of rules—namely, hard constraints—define the valid and invalid data variable-to-encoding mappings; Hu et al. [2] give an example of an invalid mapping as encoding a categorical variable with the y-position of a line chart. The other class of rules defines which encoding, of the valid candidates, should be preferred for some data variable. An example of such rules implemented in Voyager [8] is that, for a quantitative variable, the x or y-position encoding should be preferred over point encodings, and point encoding should be preferred over text encoding. These latter rules are intended to optimise certain properties of visualisations, such as perceptual effectiveness [7], [8]. The preferences for one encoding over another are derived from the literature on visualisation principles.

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.

# Methodology

We divide this work into the following sub-problems.

* What are the principles of useful and communicative geospatial visualisations?
* How can the identified principles be implemented as a program that can rank geospatial visualisation effectiveness?
* How can the set of possible valid geospatial visualisations be efficiently enumerated?
* How can a rule-based geospatial visualisation recommendation system be developed using the ranking and enumeration methods developed, and how can it be tested?

## What are the principles of creating effective geospatial visualisations?

Our initial step will involve identifying a small number of principles and guidelines for effective geospatial visualization. We will accomplish this by reviewing relevant literature such as Cleveland and McGill’s work [13]. As part of this review of the literature, we will also identify visual encodings that are relevant for geospatial visualisations.

Besides reviewing the literature, we will experiment with different identified visual encodings on geospatial datasets to identify obvious principles of geospatial data visualisation. These experiments will be done using Python on a Jupiter Notebook, and the geospatial datasets will be sourced from open data portals such as The Africa GeoPortal [14].

## How can the identified principles be implemented as a program that can rank geospatial visualisation effectiveness?

We will give each visual encoding and data type pair a rank based on the visual encoding’s effectiveness at visualising the data type. The ranking will be based on the visualisation principles identified. A visualisation will then be ranked by aggregating the ranks of its constituent visual encodings.

## How can the set of possible valid geospatial visualisations be enumerated?

Following ATP [7] and Voyager [8], we will enumerate the set of possible geospatial visualisations for a given dataset by first composing permutations of the dataset’s variables and the identified visual encodings. This will yield a set of visual encoding and data variable pairs, for example, crop type and shape, crop type and colour hue, crop yield and colour hue, and crop yield and point size. Further, a permutation of these pairs will be composed, where each element in the permutation will contain up to three variable-encoding pairs. This will be the set of possible geospatial visualisation designs for the given dataset. This set will then be filtered to remove invalid designs, for instance, a design containing both crop type and colour hue, and crop yield and colour hue is invalid and will be filtered out. This will yield a set of valid geospatial designs for the given dataset.

## How can a rule-based geospatial visualisation recommendation system be developed using the ranking and enumeration methods developed, and how can it be tested?

We will combine the developed geospatial visualisation ranking and enumeration functions to form a geospatial visualisation recommendation system. Given a dataset, the system will enumerate the dataset’s valid geospatial visualisations, rank them and then output a set number of the highest-ranking ones. This system will be tested with the help of an individual knowledgeable in visualising geospatial data. We will evaluate the system by assessing the number of the system’s recommended visualisations that the knowledgeable individual deems useful and communicative, and the number he deems not useful and communicative. The higher the number of useful and communicative recommendations the system gives, the higher we will rate its performance, and the higher the number of recommendations that are not useful and communicative the system gives, the lower we will rate its performance.

The visualisation recommendation process is depicted in Fig. 1.

Fig. 1. The visualisation recommendation process

Dataset

Data variable-Encoding Matching

Variable-Encodings Permutation

Filter

Ranking

Variable Encodings

Visualisation Designs

Valid Designs

Visualisation Recommendations

##### 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] 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.

[3] K. Stock and H. Guesgen, ‘Chapter 10 - Geospatial Reasoning With Open Data’, in Automating Open Source Intelligence, R. Layton and P. A. Watters, Eds., Boston: Syngress, 2016, pp. 171–204. doi: https://doi.org/10.1016/B978-0-12-802916-9.00010-5.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] J. Bertin, Semiology of graphics : diagrams, networks, maps. Madison, Wis: University of Wisconsin Press, 1983.

[13] 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.

[14] ‘Africa GeoPortal - powered by Esri’. Accessed: Apr. 25, 2024. [Online]. Available: https://www.africageoportal.com/