Simulated Annealing for Geospatial Visualization 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 effectiveness of simulated annealing in recommending effective geospatial data visualizations. An effective data visualization is one that accurately represents useful patterns in data and efficiently communicates these 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].

A. 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.

B. Research Question

Given the combinatorial nature of visualization design decisions, it is probable that the simulated annealing technique (defined in section II) is a good fit for the problem of finding a set of design decision combinations that produce effective geospatial visualizations. We, therefore, seek to answer the question, *how accurate is SA at finding useful and communicative visualizations for given geospatial datasets?* Accuracy is defined as the fraction of useful and communicative visualizations identified out of those identified by a knowledgeable human.

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 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.

D. Scope

This work will only consider enough design decisions to test the effectiveness of simulated annealing in finding optimal design decision combinations. We will not attempt to develop a comprehensive system considering all known design decisions for geospatial visualization, rather than attempting to build a comprehensive system that encompasses all known design decisions. By adopting this approach, we aim to streamline the experimentation process and provide insights into the capability of SA in navigating the design space efficiently.

II. Importance and Prior Work

A. Prior Work

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

The rule-based systems are exemplified by Mackinlay's A Presentation Tool (APT) [4]. The development of APT involved codifying visualization design criteria retrieved from Bertin's Semiology of Graphics [5], a seminal work on visualization. These criteria specifically measured the expressiveness and effectiveness of data visualizations. The codified criteria were used to computationally identify good visualizations from a sample of possible visualizations. This,

along with subsequent works [6], demonstrates the feasibility of representing data visualization principles and guidelines as programmable quantitative measures.

The supervised machine learning-based systems [2], [7], [8] involve learning a function that maps datasets to effective visualizations 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 knowledgeable people who had spent a lot of time visualizing data. Despite the superior performance of these systems, a significant amount of effort and cost is required to acquire and prepare the training data.

Notably, all prior works identified, consider only typical tabular data and for geospatial visualization, rather than attempting to build a comprehensive system that encompasses all known design decisions. By adopting this approach, we aim to streamline the experimentation process and provide insights into the capability of SA in navigating the design space efficiently. 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.

B. Simulated Annealing

Mackinlay [4] demonstrates that even simple datasets can have billions of design choice combinations. An increase in dataset complexity leads to a combinatorial explosion of possible visualization design choices. This makes finding optimal or effective visualizations using traditional computation methods impractical.

Simulated annealing (SA) [9] is a widely used algorithm for solving combinatorial optimisation problems. It works by intelligently sampling from a large solution space. It uses a specified objective function to compare the "goodness" of the sampled elements and gradually narrows down to an optimal solution. SA has been applied in complex combinatorial problems such as VLSI design optimization [10], the travelling salesman problem with thousands of cities [9] and "painting-from-polygons", involving finding a combination of polygon shapes and arrangements that best approximate a painting [11]. SA is a computationally efficient variant of genetic algorithms. Nonetheless, it has been shown that it can perform comparably to the classic genetic algorithms [12]. In contrast to supervised machine learning methods that learn an input/output mapping function from labelled data, SA searches for an optimal solution within a given solution space using a given objective function. This eliminates the need for collecting and preparing training data. However, defining the objective function can pose a challenge. A flow chart of SA is shown in Fig. 1.

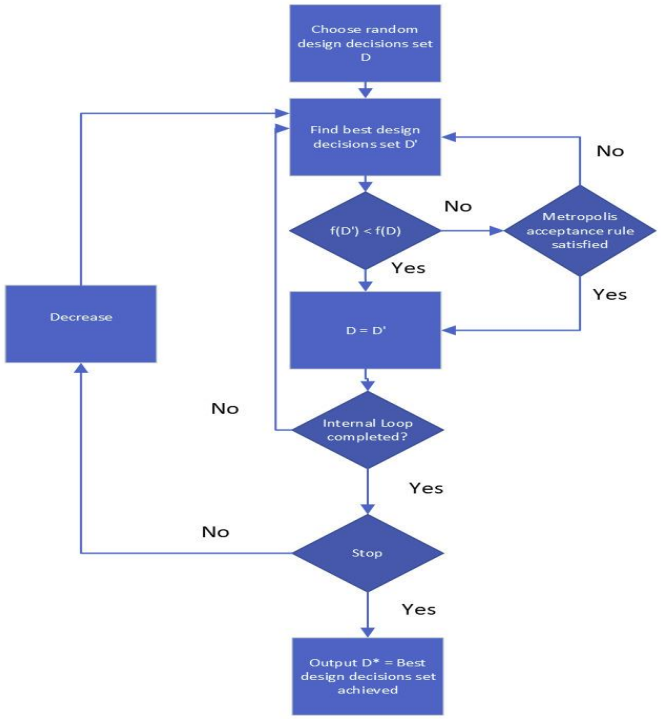

Fig. 1. Flow chart of simulated annealing

III. METHODOLOGY

We divide this work into the following sub-problems.

- What are the principles of creating effective geospatial visualizations?
- What design choices are involved in creating geospatial visualizations, and how do these choices impact the adherence to the identified visualization principles?
- How can the principles of effective geospatial visualization be represented as objective functions for a simulated annealing algorithm?
- How can we develop a simulated annealing model that recommends effective geospatial visualizations and test its performance?
- What simplified real-world problem can be used to test the resulting model's performance?

A. What are the principles of creating effective geospatial vizualizations?

Our initial step will involve identifying principles and guidelines for effective geospatial visualizations. We will accomplish this by reviewing relevant literature. Mackinlay et al. [6] recommend Bertin's "Semiology of Graphics: Diagrams, Networks, Maps" [5], Tufte's "Beautiful Evidence" [13] and Few's "Show Me the Numbers" [14] as good sources of general visualization knowledge.

B. What design choices are involved in creating geospatial visualizations, and how do these choices impact the adherence to the identified visualization principles?

Secondly, we will identify the design choices involved in creating geospatial visualizations and determine how they affect the observance of the identified principles of visualization. Design choices might include decisions about the type of map to use or which combination of columns in the dataset to consider. This step will involve a review of the relevant literature.

C. How can the principles of effective geospatial visualization be represented as objective functions for a simulated annealing algorithm?

We will then attempt to represent the identified principles and guidelines as measurable metrics and objective functions. For example, if one guideline says that a visualization's elements should not overlap to the point of obscuring its message, we will identify a way to measure this overlap given a set of design choices (like where to place labels). This measure will then be included in an objective function aimed at minimizing message-obscuring overlap.

D. How can we develop a simulated annealing model that recommends effective geospatial visualizations?

We will then model the problem of recommending geospatial visualizations using simulated annealing. Each step of the simulated annealing iteration will choose a combination of the identified design choices and assess the resulting geospatial visualizations using the previously developed objective functions. The output of the model will be a set of combinations of design choices the simulated annealing identifies as most effective for a given dataset.

E. What simplified real-world problem can be used to test the resulting model's performance?

We will use vehicle tracking data to test the model's performance. This data contains frequently recorded location data for a fleet of vehicles, along with vehicle identification data. A visualization of interest to, say, a fleet manager could be route maps showing the movement of individual vehicles, and geospatial heatmaps showing the most commonly shared routes. Given this kind of data, will the developed model recommend the mentioned useful visualizations? Also, will it recommend visualizations that communicate effectively? For instance, will it recommend route maps with only a few routes, and not so many routes that the visualization becomes unreadable?

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*, (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] 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.

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

[6] 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.

[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] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, 'Optimization by Simulated Annealing,' *Science*, vol. 220, no. 4598, pp. 671–680, May 1983, doi: 10.1126/science.220.4598.671.

[10] D. F. Wong, H. W. Leong, and C. L. Liu, *Simulated Annealing for VLSI Design*, vol. 42. in The Kluwer International Series in Engineering and Computer Science, vol. 42. Boston, MA: Springer US, 1988. doi: 10.1007/978-1-4613-1677-0.

[11] R. Dahmani, S. Boogmans, A. Meijs, and D. van den Berg, 'Paintings-from-Polygons: Simulated Annealing,' Amsterdam, Sep. 2020.

[12] K. Park and B. Carter, 'On the effectiveness of genetic search in combinatorial optimization,' in *Proceedings of the 1995 ACM symposium on Applied computing*, (SAC '95). New York, NY, USA: Association for Computing Machinery, Feb. 1995, pp. 329–336. doi: 10.1145/315891.316011.

[13] E. R. Tufte, *Beautiful Evidence*, First Edition. Cheshire, Conn: Graphics Press, 2006.

[14] S. Few, *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, First Edition. Oakland, Calif: Analytics Press, 2004.