GeoGIG: A Guidance-based Visualization Framework for Spatiotemporal Data

Behrooz Omidvar-Tehrani
Department of Computer Science
The Ohio State University
omidvar-tehrani.1@osu.edu

Plácido A. Souza Neto, Gustavo Guerino
Federal Institute of Rio Grande do Norte
IFRN, Brazil
placido.neto@ifrn.edu.br, gustavo.guerino@academico.ifrn.edu.br

***Abstract*—Spatiotemporal data is becoming increasingly available in various domains such as transportation and social science. There exists huge potentials to discover patterns and trends in this data for better decision making. A visualization suite provides visual interpretations of this data. However there exist three principled challenges which are not perfectly addressed in current visualization tools: i. the tool should be "generic" enough to provide insights for any spatiotemporal data; ii. it should be "interactive" to facilitate sub-second analyst-tool loops; iii. it should be able to "guide" analysts in an exploratory analysis scenario through potential interesting directions. In this paper, we introduce GEOGIG, an interactive guidance-based visualization framework for spatiotemporal data. We discuss the functionality of GEOGIG and provide two scenarios that illustrate the usability of GEOGIG on different spatiotemporal datasets.**

I. Introduction

Nowadays, there exists huge amounts of spatiotemporal data in various fields of science. Understanding patterns and trends through visualizing spatiotemporal data improves decision making. Some instance applications of spatiotemporal data analysis are smart city management and autonomous transport. Traditionally, an exploratory analysis scenario begins by preparing the data and employing a data visualization product like Tableau[1] or Spotfire[2] to explore a subset of data. However, with the growing size of spatiotemporal datasets, this classical analysis approach is not practical anymore. We recognize following challenges in visualizing spatiotemporal data.

Genericness. Most visualization suites do not generalize to different types of data. This is a critical challenge in spatiotemporal context due to its huge diversity from transportation data to geo-political and geo-health data. Although tools like D3[3] and VEGA [4] introduce grammars for generic visualization, formalizing a problem in form of a grammar needs some expertise which an analyst may not have.

Interactivity. The main focus of current approaches is to generate a single-shot visualization. However, in most analysis scenarios, there is a need to interact with the tool by manipulating data (filter, aggregate, etc.) Due to the gigantic size of the data and inefficient infrastructure, the interaction usually takes up to minutes and impose unnecessary waiting time on the analyst. Hence there is a need for a visualization tool to be interactive.

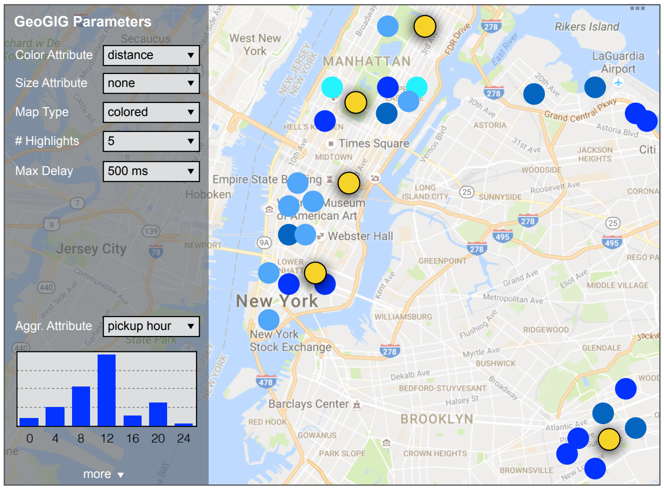
GeoGIG Parameters
Color Attribute: distance
Size Attribute: none
Map Type: colored
# Highlights: 5
Max Delay: 500 ms
Aggr. Attribute: pickup hour

Fig. 1. GEOGIG depicts few New York taxi drop-off points. Five points are highlighted by the guidance component. Pick-up hour is selected as the aggregated attribute.

Guidance. Even in an interactive visualization tool, another critical challenge is that often analysts are lost in the huge amount of visualized points and cannot easily decide "what to see next". Hence a guidance approach is necessary in order to automatically highlight few directions for further investigations.

To tackle the aforementioned challenges, we propose GEOGIG, i.e., a generic, interactive and guidance-based approach for spatiotemporal data. GEOGIG helps analysts to load and visualize their spatiotemporal data of any type without following any complicated grammar. Then the analyst observes patterns and trends from different parts of data interactively. Also, GEOGIG provides recommendations in each interactive step to guide the analyst through potential interesting directions.

II. Framework

GEOGIG employs various components and strategies to tackle the challenges of genericness, interactivity and guidance. The aim of the proposed application is to i. enable analysts visualize any spatiotemporal data with the least burden, ii. interact with the data in sub-seconds and iii. receive recommendations during the analysis process for further investigations.

GEOGIG is designed as a web service and runs in a browser. Figure 1 illustrates a screen-shot of GEOGIG on New York taxi dataset. To initialize the analysis, the analyst needs

[1] *http://www.tableau.com*
[2] *http://spotfire.tibco.com*
[3] *https://d3js.org*

to drag-and-drop her dataset (in form a CSV file) into the browser. GEOGIG exploits WEBTABLES [1] to find the best matching schema for the input dataset. GEOGIG performs the following execution command to visualize the dataset:

```
geogig ds=name [param=val]* [udf: param=val]*;
```

The only required parameter for GEOGIG is the dataset `name` which is given by the drag-and-drop. A set of visualization settings in form of `[param=val]` customizes the visualization. For instance in Figure 1, `color=distance`, hence drop-off points are colored darker if they belong to longer trips. Interestingly, points closer to airports are often darker as often longer trips are need rich airports.

GEOGIG supports the integration of spatiotemporal User-Defined Functions (i.e., `[udf: param=val]` in the execution commands) to satisfy specific needs of analysts. For instance, an analyst may create a UDF to color points in red if their tip amount is higher than $4.

GEOGIG recognizes three sets of attributes: *principle*, *aggregation* and *peripheral*. The *principle set* contains attributes which are necessary to visualize data on a geographical map, i.e., latitude and longitude. Altitude can also be considered in case of 3D projection (e.g., aviation data analysis). If GEOGIG fails to automatically complete the principle set, it will ask the analyst to manually mark those attributes. The *aggregation set* contains ordinal attributes which can be aggregated to provide complementary insights on visualized points, e.g., time-of-day, week day, etc. In Figure 1, "pick-up hour" is selected as the aggregated attribute and the abundance of points in 24 hours is shown in a histogram. Last, the remaining set of attributes is called *peripheral*. Hovering each point on the map will pop-up the list of peripheral attributes for that particular point.

A histogram exhibits aggregated counts for an attribute of choice from the set of aggregation attributes. GEOGIG exploits Crossfilter charts [5] as the histogram which enables *coordinated views*, i.e., a brush on the histogram will immediately update the map. Normally this task needs a query execution per brush, which is time-consuming. By exploiting the notion of *incremental queries*, the task evolves to sub-second execution.

GEOGIG scans the input dataset only once. For further analysis steps, it employs a *caching* strategy so that it doesn't require a connection to the dataset more than once and all other filters and brushes will be done online without re-querying. We also consider a *sampling* strategy to deliver visualizations in sub-seconds independent from the query result size. The combination of caching and sampling mechanisms assures that all analyst's requests can be served without delay independent from the nature of the request. This enables interactive analysis of spatiotemporal data.

At each step of the interactive process, a guidance component recommends a limited set of points which are potentially interesting for the analyst for further investigation. The guidance component is an adaptation from our previous work [3] to the spatiotemporal context. Figure 1 illustrates recommended points in yellow. Intuitively, the guidance component returns k points which are *highly relevant* to previous analyst's choices. Those k points are also *highly diverse* to cover the analysis space so that the analyst can investigate on different aspects of data. The guidance component is an optimization algorithm which aims to maximize relevance and diversity. However, to respect interactivity, a time limit will be given so that the algorithm performs a best-effort strategy up to the limit. The value of time limit and the size constraint (k) can be set in the setting bar (Figure 1 left).

III. DEMONSTRATION

Data. We illustrate the functionality of GEOGIG in form of realistic scenarios on New York taxi[4] and New York bike[5] datasets. We chose those datasets because they are publicly available and frequently exploited in spatiotemporal research (e.g., in [2]). However, particpants can bring their own spatiotemporal datasets in form of a CSV file to apply to GEOGIG.

Implementation. The GEOGIG engine is implemented in Python 2.7.10 and the graphical user interface is based on Node.js. The visualization components are implemented using D3[6] and the geographical map is based on GOOGLE MAPS API[7]. The demo is fully client-side and on-the-browser which makes it platform-independent.

Scenario 1. We employ New York taxi dataset for this scenario. The dataset contains 173,179,759 records of taxi trips and 18 attributes such as pickup and dropoff date/time, passenger count, tip amount and trip distance. Consider Lucas, a data scientist who is tasked to optimize New York taxi trips. Focusing on cab-idle locations, he wants to discover which neighborhoods work the best for which drivers to increase the overall availability. Lucas employs GEOGIG and follows a case-by-case inspection as his analysis methodology. We show how Lucas benefits from the guidance component to discover relevant points to his interests. We also show that he observes different aspects of data without delay due to interactivity considerations.

Scenario 2. We employ New York bike dataset for this scenario. The New York bike dataset contains bike trips from 2013 to 2016, and 15 attributes such as start/end station, trip duration and distance. Consider Stella, a data scientist who is tasked to optimize New York bike station locations in New York by analyzing bike historical data. We show how guidance component helps Stella to recognize locations with similar profile in diverse locations. Also, we show that she can focus on different parts of data (e.g., male and female riders) to make use-case studies interactively.

REFERENCES

[1] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. *Proceedings of the VLDB Endowment*, 1(1):538–549, 2008.

[2] J. Freire, A. Bessa, F. Chirigati, H. T. Vo, and K. Zhao. Exploring what not to clean in urban data: A study using new york city taxi trips. *IEEE Data Eng. Bull.*, 39(2):63–77, 2016.

[3] B. Omidvar-Tehrani, S. Amer-Yahia, and A. Termier. Interactive user group analysis. In *CIKM*, pages 403–412. ACM, 2015.

[4] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer. Reactive vega: A streaming dataflow architecture for declarative interactive visualization. *IEEE transactions on visualization and computer graphics*, 22(1):659–668, 2016.

[5] I. Square. Crossfilter: Fast multidimensional filtering for coordinated views. 2013.

[4] *https://data.cityofnewyork.us/view/gn7m-em8n*

[5] *https://s3.amazonaws.com/tripdata/index.html*

[6] *https://d3js.org*

[7] *https://developers.google.com/maps/*