Multivariate data visualization using glyphs on a geospatial map: A case study analysis on car accident data in the US

Prerana Venugopal, Srishti Jaiswal, Sudesh Pawar, Jagannath Jayanti

Abstract— Geospatial data is generated on a large scale daily, illustrating the diversity of data attribution across various geospatial territories such as counties, states, countries, and so on. The data is extremely sparse and large, and each geographical point defined by latitude and longitude has several dimensions and attributes associated with it. Given the number of possible correlations between the attributes, the data exploration procedure in such a case is quite challenging. The number of dimensions in the data increases the complexity. As a result, to aid in this type of data visualization, we have created a dashboard that displays multivariate data on a choropleth using several types of glyphs. The primary purpose of this study is to determine the number of visual encodings that can be placed on a glyph to aid in the sensemaking process. To do this, we conducted a case study using car accident data from the United States, and parameters such as time of day, accident severity, and weather condition are represented as glyphs on the geospatial map. The case study intends to help the corresponding government authorities identify the reasons that cause accidents, allowing them to gain insights into how to prevent such situations in the future. The evaluation study also offers details about the encodings that are appropriate for multivariate data processing.

Index Terms—Choropleth, geospatial , multivariate data, multidimensional data, glyphs, encoding, car accident, case study

1 Introduction

The type of representation of data that has geographical factors in it is known as geospatial data. Visualizing such kind of geospatial data is known as geospatial visualization. Such visualizations are of great interest to a broad audience for several purposes. For example, data visualizations about a particular region's weather can help metrologists make an informed decision about forecasting. However, most current visualization techniques focus on representing each factor of such data using single-dimensional shapes called glyphs. These glyphs, sometimes called icons, are graphic objects that use size, shape, and colour to represent one or more data values. Such glyphs can be as simple as a circle or a square to something as complex as a bar chart or a sunburst chart. However, as discussed before, most of the current geospatial systems still do not fully utilize all dimensions of the glyphs, as each of the glyphs represents only one geospatial data entity. In connection to a geospatial data variable, choropleth maps show separated geographic areas or regions that are coloured, shaded, or patterned. This offers a method for visualizing data throughout a region, which might reveal variations or trends within the space depicted. The data variable represents itself in each map area using a colour progression. Typically, this might be a transition from one colour to another, a progression of just one hue, from transparent to opaque, from bright to dark, or the full gamut of colours. This research work focuses on the ability to combine these two representations (choropleth maps and multi-dimensional glyphs) into a single representation by exploring the placement of multi-dimensional glyphs on a choropleth map where each dimension of the glyph, such as size, colour, shape, motion will represent a different entity of the geospatial dataset. Such representation will help the users to correlate multiple factors related to the exact geographical location with the same glyph, thus eliminating the requirement to visualize multiple choropleth maps and observe the correlations manually. However, when using multi-dimensional glyphs, there must be a limit to the number of dimensions. It is logical that as the number of dimensions increases, the complexity of the representation of such glyphs increases with it too. Therefore, the core objective of this study is to find this limit and the set of encodings that are more visually sensemaking than others. To achieve this objective, we first explored the current work in this research field, formed our design thesis based on this conducted survey, then implemented the proposed thesis on a data problem involving US Accidents data as a case study and performed a user evaluation to identify the limit on dimensions and set of encodings. The data contain attributes such as time of accident, severity, weather condition, and the year of accident occurrence plotted against the frequency of accidents occurring in the corresponding state. This is integrated on a visualization dashboard which uses encodings such shape, colour, size, motion and angle and position.

2 Related works

The challenge of placing multi-dimensional glyphs on a choropleth map has yet to be explored significantly, but some existing work exists in the field. Before we finalized our approach, we surveyed some of this research work to help us understand. Chen et al. [1] have offered, in their work, an overview of the commonalities across various research that concentrate on spatiotemporal data. It offers a fresh way of looking at the various methods that make up the cutting-edge methodology for data investigation. Lie, and the co-authors [2] of this study describe the critical elements of a glyph-based visualization design. The paper's primary goal is to assist readers in achieving successful outcomes by using a suitable framework while creating visualizations. In order to properly use glyphs in the design space and design channels that are currently available geometrics, optical, topological, and semantic channels—the paper [3] presents a design criterion. It also attempts to address the most common queries when designing glyph-based visualizations. Ward et al. [4] provide a thorough description of how glyphs are generally employed to be represented in a visualization. The many layouts that may be utilized to put the glyphs and how to do so are thoroughly discussed in the paper. In this research work [5], we offer a solution to overcome the cluttering issue by dividing the set of all objects into smaller subsets, each visually represented by an aggregated multivariate glyph showing the distribution of its objects' direction, type, and velocity. The authors of this paper [6] have suggested utilizing procedural modeling to address the Geographical Visualization modeling issue. This multi-layered structure enables the model to execute queries in the context of glyph intersection, resolve any potential glyph collisions, and map abstract data to a geographical domain. Chung et.al.[7] offer a conceptual framework based on glyphs for interactive multivariate data sorting as a step in the visualization process. In their research, Kim et al. [8] focused on encoding multivariate data in a 2D space using bristlecone glyphs on the map. This study [9] explains the glyph-generating process and discusses several perceptual problems that can distinguish between effective and inefficient glyphs. The primary objective of this study [10] was to add 3D structures to

Fig 1: Dashboard displaying multiple dimensions using glyph like pie chart, bar chart and colour gradient.

glyphs, such as curvatures, to improve the effectiveness of user interpretation of multivariate data. Netek et.al.[11] seek to thoroughly investigate the user preferences for traffic accident data-based heat map visualization of several categories. The experiment's goals were to lay out critical rules and to illustrate the expected benefits and drawbacks of heat map usage. The primary goal of this research work [12] is to develop a program called "VTGeo" that will enable users to see geospatial data on various maps, including heat maps, scatter plots, and road maps. When users interact with the database, the authors have devised a way to translate the data structures they use or construct into geographical information. To investigate how visualizations could assist in detecting the effects of two specific traffic rules, the research piece [13] contrasts GeoMultiVis with other visualization tools. According to the idea of the visualization pipeline, the article [14] suggests breaking the process of glyph-based 3D visualization into three discrete processes, namely Data Mapping, Glyph Instantiation, and Rendering. An important conclusion from this current study is that having the proper glyph design is essential for successful glyph-based visualization. This survey [15] aims to understand better how the star glyph-based method is applied while working with multivariate data. In their article, McNabb et al. [16] detail the development of an algorithm pipeline for this process and demonstrate how the amount of detail in the resulting picture can be altered by the user. The technique uniquely combines guided glyph placement, degree of detail, dynamic zooming, and smooth transitions.

3 IMPLEMENTATION

The goal of the visualization is to plot multiple visual encodings using different visual encodings using a glyph on a choropleth. The primary layer in our visualization is the choropleth, representing a color gradient of the frequency variation across all the states. The geographical boundaries are marked clearly to encode the boundaries of each state, and this map is then color coded with the frequency of the dimension that we want to measure. The glyph layer is superimposed on the map, with each glyph representing dimensions of a particular state. The glyphs are mostly objects and can be represented using any shape and color. In our project, we considered visual elements such as circles, bar graphs, pie charts and shapes such as clouds, snow, etc. For the user interface, we have primarily considered five different views:

The initial landing page/ Overall View: This view is divided into two sections, an interactive choropleth representing the frequency of the measure distributed across different states and an interactive space where the user can choose the kind of visual encoding he is interested in visualizing. Each chart has a different number of visual encodings embedded into the choropleth, allowing the user to choose his suitable visualization design.

State-wise Statistics View: Once the user clicks on a particular state, a hyperlink called “Get Statewise Stats” appears at the bottom end, and on click of this hyperlink, a new view appears that provides the comprehensive statistics report of each state using a bar graph and word cloud. The user can also select more than one state or analyze multiple states by region, and the statistics report gets updated dynamically.

Glyph View: Once the user selects his preferred visual encoding, the landing page is navigated to a glyph view representing dimensional encoding using glyphs. We have considered five primary visual encodings for glyphs: colour, size, length, shape, and area.

Time Series View: To illustrate motion encoding over time, we have a time series view representing how each factor annotates at different time instances. The primary encodings are size, area, position, and time.

The synchronization view outlines how multiple factors contribute to the distribution at any time. All the time series charts are visualized at once and thus give the user the flexibility to correlate and understand how one affects the other.
The approach to implementation consisted of three steps:
Data Exploratory Analysis: The initial raw data were preprocessed to clear any null values, and convenient features were cleaned and extracted into a .csv format. This entire preprocessing was done through Python using NumPy and Pandas library.

Prototype and initial visualization design: The initial data exploratory analysis revealed factors that could influence the final visualization. A basic choropleth was created, and the different glyphs were plotted on this map using the Bokeh library in Python to understand the glyph selection and placement strategies across different states.

Final dashboard creation: From the inferences obtained in the initial prototype, the visualization dashboard was designed using the tableau framework. The cleaned data was imported as tables in the Tableau workbook, and multiple visual elements were created using the segregated data. This workbook was hosted on a Tableau public server and deployed as an HTML file.

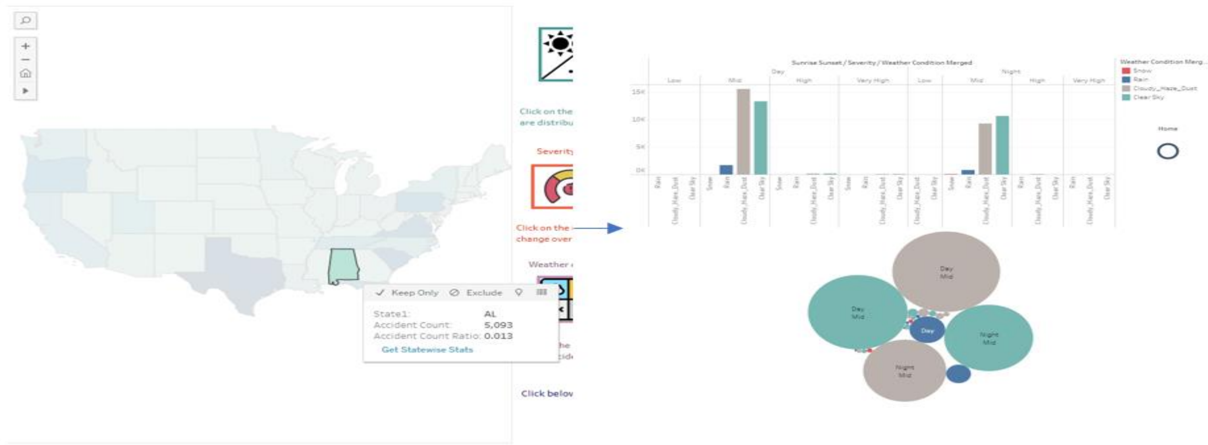

Fig 2: Dashboard displaying state-wise statistics of different factors for each state.

The user interaction involves multiple interactivity tools such as selection, highlight, details on demand, linking, and brushing to navigate between different views in the dashboard, thus abiding by Schneiderman's mantra. As per McGill and Cleveland's ranking, the position is giving the first preference to place glyphs on the map. The other significant visual encoding used is the color to denote the gradient in the choropleth. Apart from this, length encoding is used in bar graphs, area and angle are used in pie charts, size is used to denote the frequency of the attribute, and motion is used to represent time in the time series chart.

4 Case Study – Visualizing Traffic Accidents Using a Choropleth

4.1 Data

For the purpose of exploring placement of multi-dimensional glyphs onto a choropleth map as part of the visualization study, we wanted to use an appropriate problem as a case study to showcase the implemented work. After careful brainstorming about the potential use cases of our proposed work, we decided to work on visualization of accidents in United States from 2016 to 2021. For this purpose, we explored several websites and articles to find raw data. Many existing authors have already collected data from different sources for similar research works. For our case study, we felt that the data used by Moosavi et.al. [17] in their research work was most appropriate and since it was open sourced via Kaggle, we decided to use it. The original dataset consists of 40 factors and over a million accidents, however it is very unclean and for many tuples, some of the entries are missing. Therefore, before proceeding with the visualization work, we used Python along with libraries such as Pandas to clean the data and performed some statistical measures on the data and found out that mainly there are only 3-5 factors which contribute and hence, we filtered the data to include the features with highest correlation. A data frame was created using this information to hold state-wise information on each of the corresponding factors. Since the population of each state misleads the frequency of accident occurrence, the entire data was normalized, and this ratio of accidents per capita was plotted on the choropleth to gain valuable insights. This normalization processing was carried out on all measures such as count and frequency.

4.2 Implementation Design

The basic element of our visualization design is a choropleth which represents the ratio of accidents per capita distributed across different steps. This is visualized on a landing page which acts as the starting point for our visualization dashboard. On the left part of the landing page, we have an interaction panel that is divided into four subsections based on the number of dimensions that the user wants to visualize.

The first subsection focuses on visualizing ratio of day and night accidents using various visual encodings. The single dimension encoding represents the frequency of day v/s night accidents in two different choropleth maps. The colour gradient indicates the higher accident ratio in each state. The background colour using bright and dark represents the day and night accidents respectively. The two-dimensional comparison view using bar graph can be viewed by clicking on the bar chart icon on the landing page. The night accidents are represented through dark blue, and the day accidents are represented through orange colour on the bar chart. On hovering on a state, a pop menu appears that displays the information of different distribution and the state information that is currently being visualized. The third type of visualization is a pie chart that focuses on plotting all categories, one for each state. The size of the pie chart changes according to the respective accident frequency of each state. The landing page also contains a time series chart to visualize the distribution of day and night accidents over time. The primary encoding in focus here is motion and an animation video is played to visualize this. The visualization elements remain constant for the other factors such as weather and severity condition. The severity feature is divided into 4 subcategories such as low, mid, high, very high and represented using a red colour palette. The weather conditions are divided into four categories such as snow, rain, cloudy and clear sky and is represented through red, blue, grey, and dark grey color encodings. The landing page contains state-wise statistics report for all the above factors. This view can be enabled by clicking on any state in the choropleth and selecting the hyperlink. This view shows a word cloud and a bar chart that shows the major factors contributing to accidents in a particular state.

The user can zoom on any state or region of the map to see the glyphs clearly and can infer more information on the color encodings used. Brushing and linking features are also activated in the state-wise view and this gets dynamically updated on selecting multiple states. On hovering any state, a pop-up menu appears providing details of the respective state information thus adhering to details on demand interaction.

Our visualization design can currently represent four visual encodings at once in the time series view which represents motion, position, size, and color. On evaluation of our dashboard, it was found that the users preferred views that contained only two encodings rather than multidimensional representation. This can be used as a guide to design multivariate visualizations on a choropleth in the future. The current case study was done on a single dataset covering a single domain, however more study needs to be done to understand how visual encodings can vary across different domains.

5 EVALUATION

The goal of our evaluation was to explore how useful multi-dimensional glyphs were in allowing users to analyze the data. As discussed before, as proof of concept of our work, we have developed an analytical dashboard to visualize accident data where we have incorporated multi-dimensional glyphs on the choropleth map.

5.1 Participants

We performed a user study as the evaluation method for our research work. A total of 36 participants volunteered to participate in the user study. 30 of them were associated with University of Florida and 6 are based in India. The age of the participants ranged from 22 to 30 years. 8 of the participants had previous experience of working in the Data Analytics field, however none of them had any experience dealing with multi-dimensional glyphs. All the participants were asked about their proficiency in dealing with visualizations and whether they had any idea about what choropleths are.

5.2 Tasks

We studied the efficiency of visualizations by asking the participants to perform three different tasks with the help of our developed system. At the start of the session, the participants were briefed about the visualization and purpose of study, they were also given a short walkthrough about how to use the dashboard. The first task involved identifying the state with the highest number of accidents per capita during day time. For this task, the users were expected to look at the dimensions representing population and categorical distribution of time of the accident. For the second task, the participants were asked to identify the month in which Texas had highest number of accidents with very high severity. To complete this task, the users were supposed to observe the combination of motion (to identify time) and color (to identify no. of accidents classified with severity as very high). The third task , participants were asked to identify which weather conditions are considered to be of a higher risk to travel in the state of Florida. To complete this task, the users were expected to select the bar chart view and observe combination of length and color to correctly identify the answer.

5.3 Results

The users were asked to fill in a short questionnaire after completing the three tasks where they were asked to type the answers for the three tasks. The participants were also asked to select the different encodings they observed in the dashboard, difficulty in analyzing, whether the developed system was user-friendly and select the combination they felt most useful to complete the given tasks. After analyzing the questionnaire filled by the participants, we found that all the participants were able to solve Task 1 and Task 3 efficiently and for Task 2, 33 of the 36 performed it accurately. Most of the participants were able to identify the encodings of length(size), color, motion, shape. 33 participants found the combination of color and size to be most useful and 27 participants found the combination of color and motion to be most helpful. 34 of the 36 participants found use of multi-dimensional glyphs to analyze as an easy to medium difficulty task where correlation required some effort but gathering information was easy. All participants, except one found the User Interface of the dashboard to be user-friendly. Few of the participants also gave additional feedback about which aspects of multidimensional glyphs they found more useful. Overall, based on the evaluation, we concluded that the addition of multi-dimensionality in glyphs positively affected the visualization.

6 CONCLUSION

This project's goal is to help users in comprehending, interpreting, and using data glyphs for multi-dimensional data while also attempting to fit as many encodings onto the glyph as possible to avoid making the representation appear crowded. Considering this, we presented our project as a collection of interactive glyphs plotted over a choropleth map that is meant to be utilized by both expert and novice users to gain a better knowledge of the usage of glyphs for data visualization.

The main contributions of this paper are: (1) An interactive multidimensional glyph representation for multivariate data aids in comprehending the alleged cause of an accident (such as severity, weather, day/night, etc.) for the US Car Accident case study presented in this paper. (2) Accomplished placing the glyphs on the choropleth in a way that allows users to gather insights. For more targeted research, glyphs were employed to indicate many elements individually, such as the weather, severity, etc. (3) Achieved an interactive technique that incorporates coordinated multiple views and the brushing and linking approach to facilitate visual analysis.
As for the limitations, the first is that we were unable to implement more than 4 encodings for the glyph design that was presented in this paper. Our visualization hasn't been validated utilizing numerical data variables because the dataset we utilized for our case study only contained categorical data.

Currently, we are analyzing how interactive glyph visualizations might help users in understanding the factors that might affect accidents in the US and how these factors can help identify potential accident-prone zones in a case study utilizing our glyph visualization together with other visualization tools. In the coming stages, we want to get deeper into how interactive glyph visualization affects decision-making by using it periodically for daily analytics chores and monitoring various metrics like performance, scalability, and accessibility, for example.

7 ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of our peers at University of Florida and our friends back in India for helping us in the user evaluation of our proposed solution by agreeing to participate in the study. We are also grateful to the Department of Computer & Information Science, University of Florida for providing the technical facilities to perform our research work. Special thanks to Prof. Eric Ragan and Ms. Ferby Cremer for their able guidance and regular feedback which helped us in development of this work.

8 ADDITIONAL LINKS

Link to Tableau dashboard: Link to Dashboard

Link to Video dashboard: Link to Video

9 REFERENCES

[1] Chen, X., Shen, L., Sha, Z., Liu, R., Chen, S., Ji, G., & Tan, C. (2019). A survey of multi-space techniques in Spatio-temporal simulation data visualization. Visual Informatics, 3(3), 129-139.

[2] Lie, A. E., Kehrer, J., & Hauser, H. (2009, April). Critical design and realization aspects of glyph-based 3D data visualization. In Proceedings of the 25th Spring Conference on Computer Graphics (pp. 19-26)

[3] Borgo, R., Kehrer, J., Chung, D. H., Maguire, E., Laramee, R. S., Hauser, H., ... & Chen, M. (2013, May). Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. In Eurographics (State of the Art Reports) (pp. 39-63)

[4] Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization. Information Visualization, 1(3-4), 194-210.

[5] R. Scheepens, H. v. d. Wetering and J. J. v. Wijk," Non-overlapping Aggregated Multivariate Glyphs for Moving Objects," 2014 IEEE Pacific Visualization Symposium, 2014, pp. 17-24, DOI: 10.1109/PacificVis.2014.13.

[6] Karnick, Pushpak Jeschke, Stefan Cline, David Razdan, Anshuman Wentz, Elizabeth Wonka, Peter. (2009). A Shape Grammar for Developing Glyph-based Visualizations. Comput. Graph. Forum. 28. 2176-2188. 10.1111/j.1467-8659.2009.01428.x.

[7] Chung, David HS, Philip A. Legg, Matthew L. Parry, Rhodri Bown, Iwan W. Griffiths, Robert S. Laramee, and Min Chen. "Glyph sorting: Interactive visualization for multidimensional data." Information Visualization 14, no. 1 (2015): 76-90.

[8] S. Kim, R. Maciejewski, A. Malik, Y. Jang, D. S. Ebert and T. Isenberg," Bristle Maps: A Multivariate Abstraction Technique for Geovisualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 9, pp. 1438-1454, Sept. 2013

[9] Ward, Matthew & Chen, Chun-houh & Härdle, Wolfgang Karl & Unwin, Antony. (2008). Multivariate Data Glyphs: Principles and Practice. 10.1007/978-3-540-33037-0_8.

[10] Forsell, Camilla & Seipel, Stefan & Lind, Mats. (2006). Surface Glyphs for Efficient Visualization of Spatial Multivariate Data. Information Visualization. 5. 112-124. 10.1057/palgrave.ivs.9500119.

[11] Netek, Rostislav, Pour, Tomas and Slezakova, Renata. "Implementation of Heat Maps in Geographical Information System – Exploratory Study on Traffic Accident Data" Open Geosciences, vol. 10, no. 1, 2018, pp. 367-384.doi.10.1515/geo-2018-0029.

[12] G. Shengnan, X. Jianqiu and S. Tao, "VTGeo: A Visualization Tool for Geospatial Data," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020, pp. 2361-2364, doi: 10.1109/ICMCCE51767.2020.00510.

[13] Wagner Signoretti, Isabel Manssour, and Milene Silveira (AMCIS 2019 Proceedings), GeoMultiVis: helping decision-making through Interactive Visualizations from Geospatial Multivariate Data [2019]

[14] Andreas E. Lie, Johannes Kehrer, and Helwig Hauser (SCCG '09: Proceedings of the 25th Spring Conference on Computer Graphics), Critical Design and Realization Aspects of Glyph-based 3D Data Visualization [January 2009]

[15] Izyan Izzati Kamsani, Norharyati Md Ariff, Noor Elaiza Abd Khalid (Universiti Teknologi Mara (Shah Alam), Department of Computer Science, Faculty of Computer and Mathematical Science, 40000, Shah Alam, Selangor, Malaysia), Survey of Star Glyph-Based Visualization Technique for Multivariate Data [April 2011]

[16] Liam McNabb and Robert S. Laramee (Visual and Interactive Computing Group, Department of Computer Science, Swansea University, Bay Campus, Swansea SA1 8EN, UK), Multivariate Maps—A GlyphPlacement Algorithm to Support Multivariate Geospatial Visualization [September 2019]

[17] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. 2019. Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19). Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/3347146.3359078