LEARNING CULTURAL HERITAGE DATA IDIOSYNCRASY THROUGH SCIENTIFIC GRAPHS

L. Targa, C. Portalés, J. Sebastián, I. Coma, P. Morillo, M. Fernández

Universitat de Valencia (SPAIN)

Abstract

Data visualization provides users with intuitive means to interactively explore and analyze massive datasets, which can be dynamic, noisy and heterogeneous, enabling them to effectively identify interesting patterns, infer correlations and causalities, and support sense-making activities, making it possible to amplify human cognition.

Cultural Heritage (CH) institutions are a great source of high-quality data, that is not really being tapped into by data scientists. It is important to understand, however, that cultural data are different from those provided by natural sciences. They are seldom discreet and univocal. Regarding location and time, they are also heterogeneous, given with different granularities, and, for some cases, present uncertainty.

The vision of ClioViz is to take advantage of current technologies to extract the best possible information from huge datasets and synthetize them in visual forms that combine scientific visualizations and interactive maps. Data analysis will be used to retrieve meaningful information from datasets in the field of CH, with special focus on the visualization of the multispace and multitime complex variables, benefiting from visualization strategies coming from GIS and scientific visualizations.

Within the scope of the ClioViz project, we have developed a gallery of graphs (e.g., barplots, violinplots, spiderplots) that aims to offer visualization of the uncertainty of location for some heritage objects. This paper explores the understanding of such visualizations from a case study with undergraduate students: on the one hand, students in the field of data science; on the other hand, students in the field of art history. To that end, we have conducted a pair of workshops where the aims and scope of the ClioViz project were explained to the students, and then they individually fulfilled an online questionnaire to evaluate the understanding and/or usefulness of the proposed visualizations. Results showed that most of the students understood more straightforward visualizations such as heatmaps, scatterplots or barplots and considered them as optimal to explain the uncertainty of locations. However, more complex visualizations (violins, spiders, etc.) were considered confusing, by mainly the art history students. Students also proposed ways to improve the understanding of visualizations, such as changing some of the aesthetics and adding interaction to the graphs. Based on their feedback, the graphs have been improved.

Keywords: Data visualization; workshop; multidisciplinary, cultural heritage

1 INTRODUCTION

Cultural Heritage (CH) institutions are a great source of high-quality data, that is not really being tapped into by data scientists. It is important to understand, however, that cultural data are different from those provided by natural sciences. They are seldom discreet and univocal. Regarding location and time, they are also heterogeneous, given with different granularities, and, for some cases, present uncertainty.

The vision of ClioViz is to take advantage of current technologies to extract the most valuable information from huge datasets and synthetize it in visual forms that combine scientific visualizations and interactive maps. Data analysis will be used to retrieve meaningful information from datasets in the field of CH, with special focus on the visualization of the multispace and multitime complex variables, benefiting from visualization strategies coming from GIS and scientific visualizations.

In CH, spatial and temporal information is crucial for its conservation, documentation, and dissemination. Data is heterogeneous and comes with a certain granularity and uncertainty in both variables. These issues are summarized in Table 1, where specific examples are mentioned.

Table 1. Characteristics of Cultural Heritage (CH) data.

	Location / Space	***Time***
Heterogeneity	The location can be given by a toponym (e.g., Madrid) or by geographic coordinates (e.g., 40.4165, -3.70256)	The temporal scale can be given as textual form (e.g., eighteenth century) or numerically (18th century)
Granularity	The location of an object can be determined imprecisely, covering large regions (eg, Spain) or more specifically (e.g., Madrid)	The time variable can be indicated imprecisely (e.g., 14th century) or more specifically (e.g., year 1359)
Uncertainty / Ambiguity	The location of an object can be given by multiple locations when there is uncertainty or there is a diversity of opinions (e.g., Spain, Italy)	An object might have different times, in case of disagreement about its date (e.g., 17th, 18th centuries)

However, these characteristics (heterogeneity, granularity and/or uncertainty) can be present in both space and time. For instance, an object whose location is "Spain, Palermo" can be seen as an example of granularity (as the region it covers is a whole country) and uncertainty (since the location is given why two toponyms).

In addition, spatial and temporal data can have different meanings depending on which information they have. When talking about space, it can indicate the production place, the current location or the origin of an object. Similarly, with temporal data, a historiographic term can change its meaning (for example, Modern art is not the same as Modern Age), descriptions of the same object may reference to different periods (neogothic, neoclassical, neobyzantine, etc.) or the multiple dates that an object can contain (production, entry to the museum, restorations, changes of ownership, etc).

This paper explores the understanding of the visualizations created from a case study with undergraduate students. This study encompasses two groups: one comprising students specializing in data science, and the other consisting of students studying art history. To achieve this, we have conducted a pair of workshops where the aims and scope of the project were explained to the students, and then they individually fulfilled an online questionnaire to evaluate the understanding and/or usefulness of the proposed visualizations.

The paper is organized as follows: in the next section, we explain the methodology, that includes the workshop with the students; section 3 refers to the results, where we analyse the outcomes of a test performed during the workshop; finally, section 4 brings some conclusions.

2 METHODOLOGY

To understand which visualizations help us the most to see heterogeneity, granularity and uncertainty, a gallery of graphs has been created. These visualizations are grouped in different categories based on the examples provided at [1], and graphs in the same group display the same information. This is the key to comparing the different visualizations to determine the best use case.

The groups created shows which graphs have been used to create the galley and the main purpose of each one:

- Distribution: use to visualize the distribution of a numeric variable for one or several groups. The graphs selected in this category are violin plot and boxplot.
- Correlation: shows the relation between two variables displayed in each axe. The graphs used are scatterplot and heatmap.
- Ranking: use to compare amounts between groups. The graphs selected are barplot, spider/radar, wordcloud.
- Part of a whole: shows the proportion of each part (or parts) of a variable in relation to its total. In this case, the graphs used are treemap, donut, circular packing.
- Flow: used to compare amounts. The graph selected is Sankey.

To have a quick idea of those visualizations, Fig 1 shows icons of each visualization by group.

(a) (b) (c) (d) (e)

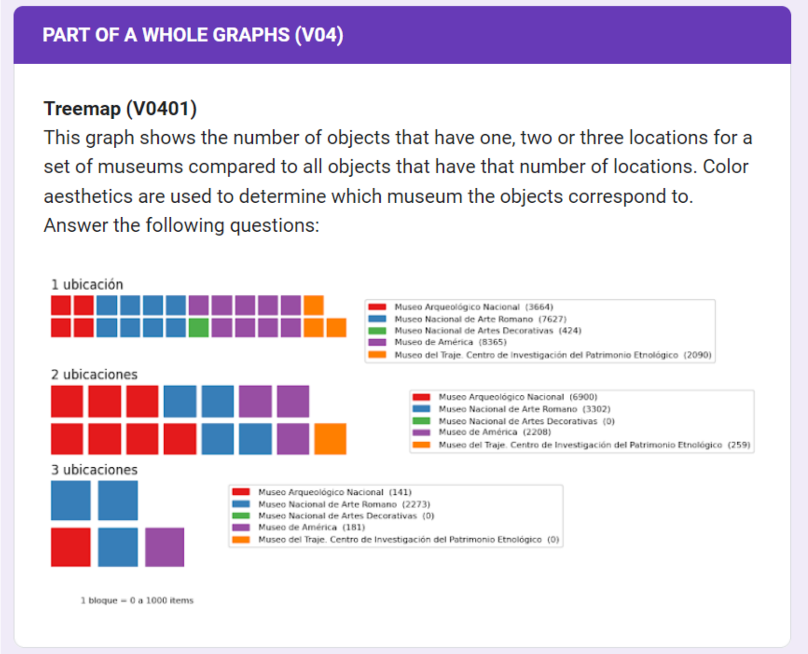
Figure 1. Examples of different graphs grouped by category, where: a) Distribution; b) Correlation; c) Ranking; d) Part of a whole; e) Flow. From [1]

To obtain feedback from the students, the first step taken was to provide an overview of the project. Since there were students from different careers (art history and data science), this explanation had to be adapted to the prior knowledge of each group. For instance, the art history students needed an introduction to how information can be accessed and how a machine could read that information and extract it from the web. Another essential part of their understanding of the project was to show them examples of how Big Data was applied in their study field. Some examples are the geoposition of events and personalities of Hispanic History [2], a spatiotemporal timeline of the British Museum [3], or some more complex dashboards like the result of the challenge with data from the National Gallery [4] and the diversity on the walls of the National Gallery of Art [5]. When explaining the project to the data science students, the explanation was focused on data cleaning and exploratory analysis, along with the steps followed to obtain the coordinates of each object.

The second part of the workshop, filling out the form, was the same for both groups. For each group of graphs and individual graph, the following questions were asked:

- This type of graph is useful for understanding uncertainty in geolocation data. Likert scale question from (1) totally disagree to (5) strongly agree with an option “I do not understand the question”.
- Can you think of another way to represent geographic uncertainty with this type of graph, or how we could improve it?
- Which graphics in this category did you like the most?

An example of the graph created, Fig 2a shows the number of objects with one, two or three locations for a set of museums compared to all objects that have that number of locations. The aesthetic of color is used to determine which museum the objects correspond to. The questions about this graph are displayed on Fig 2b; while Fig 2c presents the question related to the group.



PART OF A WHOLE GRAPHS (V04)

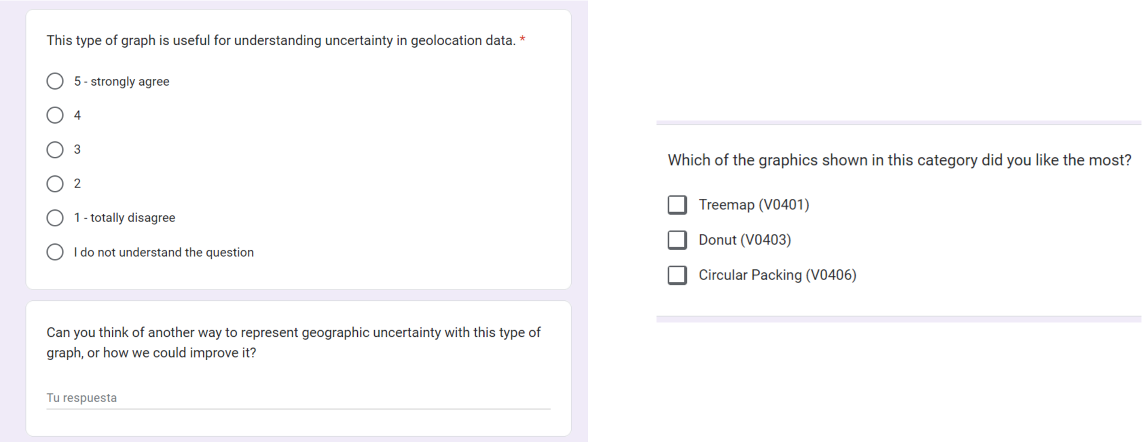
Treemap (V0401)

This graph shows the number of objects that have one, two or three locations for a set of museums compared to all objects that have that number of locations. Color aesthetics are used to determine which museum the objects correspond to.

Answer the following questions:

1 ubicación
🟥 Museo Arqueológico Nacional (3664)
🟦 Museo Nacional de Arte Romano (7627)
🟩 Museo Nacional de Artes Decorativas (424)
🟪 Museo de América (8365)
🟧 Museo del Traje. Centro de Investigación del Patrimonio Etnológico (2090)
2 ubicaciones
🟥 Museo Arqueológico Nacional (6900)
🟦 Museo Nacional de Arte Romano (3302)
🟩 Museo Nacional de Artes Decorativas (0)
🟪 Museo de América (2208)
🟧 Museo del Traje. Centro de Investigación del Patrimonio Etnológico (259)
3 ubicaciones
🟥 Museo Arqueológico Nacional (141)
🟦 Museo Nacional de Arte Romano (2273)
🟩 Museo Nacional de Artes Decorativas (0)
🟪 Museo de América (181)
🟧 Museo del Traje. Centro de Investigación del Patrimonio Etnológico (0)
1 bloque = 0 a 1000 items

(a)



This type of graph is useful for understanding uncertainty in geolocation data. *

○ 5 - strongly agree

○ 4

○ 3

○ 2

○ 1 - totally disagree

○ I do not understand the question

Can you think of another way to represent geographic uncertainty with this type of graph, or how we could improve it?

Tu respuesta

(b)

Which of the graphics shown in this category did you like the most?

□ Treemap (V0401)

□ Donut (V0403)

□ Circular Packing (V0406)

(c)

Figure 2. Questions of the form about a certain graph.

Besides the questions regarding the graphs, the form had a final section with some open questions to know the opinion of the students about the use of visualization in the cultural heritage field in helping understand the heterogeneity, granularity, and uncertainty of data. Finally, they could navigate through the gallery of graphics from which we took our inspiration so they could identify any possible graph that would be interesting for the cultural heritage field and what information they would display.

3 RESULTS

The main idea behind those workshops was to compare each group's answers. However, after reviewing all the answers, we faced a problem with the imbalance in the data: 71% of the responses obtained were from data science students and only 29% to art history students. Therefore, the next section only explores the responses from the leading group with previous data visualization knowledge.

Table 2 shows the level of understanding of each graph after looking at an example to determine if that graph was useful for understanding uncertainty in geolocation data.

Table 2. Understanding of each type of graph.

	Minimum	***Average***	***Maximum***
Violin	1	3,28	5
Boxplot	1	3,31	5
Scatterplot	1	3,59	5
Heatmap	2	4,19	5
Barplot	2	4,13	5
Spider	1	2,97	5
Wordcloud	1	2,19	4
Treemap	1	2,94	5
Donut	1	3,13	5

Circular packing	1	3,91	5
Sankey	1	2,44	5

Analyzing those results, straightforward visualizations such as heatmaps, scatterplots or barplots were considered optimal to explain the uncertainty of locations. However, more complex visualizations (violins, spiders, etc.) were considered confusing. In addition, there was one student that answered with "I do not understand the question" in some graphs (treemap, circular packing and sankey).

Fig 3 breaks down the scores for 5 graphs, one per group, and allows us to determine similar conclusions as before. If there is a mayor understanding of the graph, generally there are no answers with "I do not understand the question. However, in other cases (Sankey and violin) where there is a general confusion about the graph, answers vary among all options.

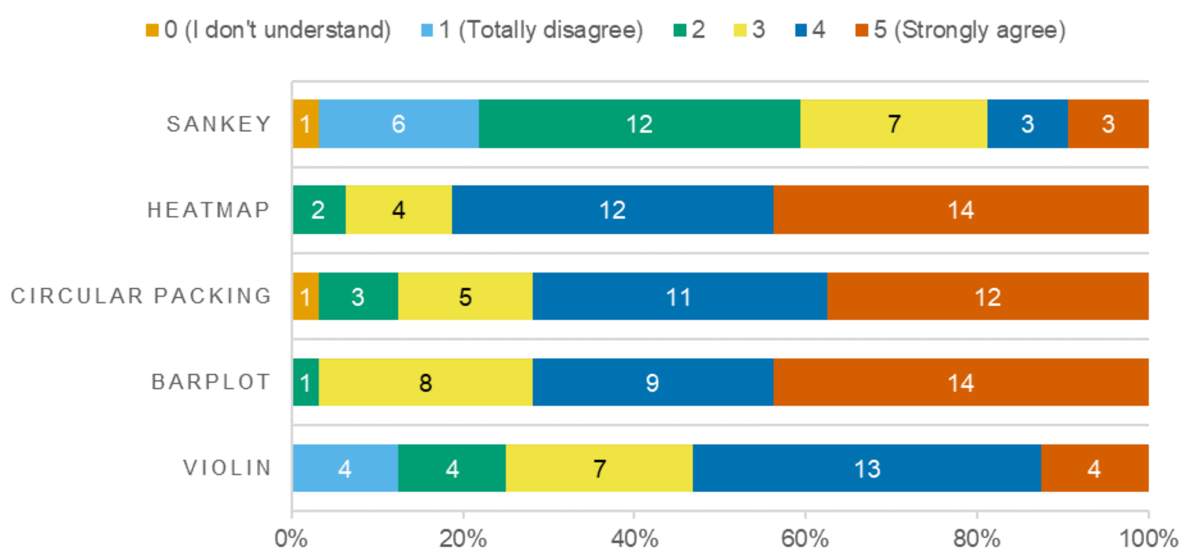


Figure 3. Number of scores (from 0 to 5) for a graph in each group

4 CONCLUSIONS

Thanks to the feedback from the students, it has been possible to expand the gallery of graph to make better examples that can explain the uncertainty in geolocation data better. Some examples involved changing the axis because students were not used to visualize a barplot with the categorical variable on the y-axis; using size as an aesthetic to visualize the amount or percentage of objects; or changing the grouping variable that could declutter the graph and to make comparisons more quickly in between groups.

Along the way, students thought that using scientific visualizations could add value to the field of cultural heritage. Some highlighted that scientific visualizations could provide a better understanding of the concepts and can help to draw new conclusions, analyse the available information, and make decisions, or that they capture in a more graphic way relevant data.

ACKNOWLEDGEMENTS

The research leading to these results is in the frame of the project "ClioViz: Visualización avanzada de datos históricos a través de mapas multidimensionales y gráficos interactivos", which has received funding from the Spanish Ministry of Science and Innovation, with the ID: PID2021-126777NB-I00. Cristina Portalés is supported by the Spanish government postdoctoral grant Ramón y Cajal, under grant No. RYC2018-025009-I.

REFERENCES

[1] Y. Holtz, “Python Graph Gallery,” *The Python Graph Gallery*. https://python-graph-gallery.com/ (accessed Sep. 06, 2023).

[2] “Historia Hispánica - Explore history through our interactive map.” https://historia-hispanica.rah.es/ (accessed Sep. 07, 2023).

[3] “The Museum of the World,” *Museum of the World*. https://britishmuseum.withgoogle.com/ (accessed Sep. 07, 2023).

[4] “GMU National Gallery of Art Data Challenge Results | Tableau Public.” https://public.tableau.com/app/profile/paul.albert2725/viz/GMUNationalGalleryofArtDataChallengeResults/StoryBoardV2 (accessed Sep. 07, 2023).

[5] “Diversity on Display: Who’s on the Wall at NGA?” https://bheggeseth.shinyapps.io/DiversityOnDisplay/ (accessed Sep. 07, 2023).