Issues and Suggestions for the Development of a Biodiversity Data Visualization Support Tool

Pawandeep Kaur [†1], Friederike Klan [‡1] and Birgitta König-Ries [§ 1]

[1]Heinz-Nixdorf Endowed Chair for Distributed Information Systems, Department of Mathematics and Computer Science, Friedrich-Schiller-Universität, Jena, Germany

Abstract

Visualizations are an important tool to transport information. However, finding the right visualization can be challenging. Using the biodiversity research domain as a showcase, we investigate where exactly these challenges are and what a tool should look like that helps scientists overcome them. Our results are based on a survey we performed.

Categories and Subject Descriptors (according to ACM CCS): A.1 [Computer Graphics]: General Literature—Introductory and Survey *(see http://www.acm.org/about/class/class/2012)*

CCS Concepts

•***Human-centered computing*** → *Scientific visualization;*

1. Introduction

To address the critical challenges of biodiversity conservation and study its impact on the ecosystem, scientists have produced a large amount of highly heterogeneous, multidimensional and distributed data. Proper visualizations are needed to decipher and comprehend the information which are inherent in this data. The key elements of successful information visualization are: choosing a visualization technique that fits the data characteristics and supports the user's information seeking goal. If these elements are ignored, people might interpret the data in an unintended way or might not understand the underlying information [KKUW07]. Visualizations in scientific articles have been criticized due to many of the following quality issues: inadequate, missing, or contradictory explanation or labeling, visual clutter and distortion, extraneous and unnecessary decoration, non-standard graphic conventions, inappropriate selection of representations (e.g., simple univariate displays when multivariate displays were needed) [CSC02] [SSS*06] [WMWG15] etc. Previous visualization usability studies [DLW*17] have investigated the reasons behind these visualization usage inadequacies and have found that users (especially scientists) lack trust in cutting-edge tools as opposed to conventional analysis mediums. They often use their own analysis and visualization tools thus do not consider the spectrum of considerations when creating visualizations.

In order to solve the problem of visualization inadequacies in scientific publications, studies [DLW*17] have provided compelling evidence that fundamental changes in the types of figures that scientists use are needed. Some of the solutions provided by them are: changing the journal policies for visualization figure acceptance, training scientists, providing more strict guidelines for graph constructions, providing rulebooks etc. Instead of training scientists in data visualization, we plan to develop a tool that assists them in the selection of a suitable visualization based on the data properties, the data domain, and the user's representational goal. In order to implement such a solution for our domain users, the foremost step for us was to understand the current visualization usage patterns, needs and aspiration from our users. To gather this information, we did a survey among biodiversity researchers to elicit direct feedback from our domain users. In this paper, we present the results from our user study to answer two broad questions: 1) which problems do our users face when selecting and producing visualizations?, 2) how should a tool that helps biodiversity researchers to overcome these issues look like? The rest of the paper goes as: information about how the study was conducted is presented in Section 2. Findings from the study and discussion are presented in Section 3. Lastly, conclusions and future directions are in Section 4.

2. Method

We performed a survey to get direct feedback from our domain users about the domain specific operations they perform with different visualizations, challenges they face in visualizing their data and the technological assistance that can support them. The results of this survey are provided in Section 3. This survey was done via the medium of a paper questionnaire and an online form at various

† www.fusion.cs.uni-jena.de/fusion/members/pawandeep-kaur/
‡ www.fusion.cs.uni-jena.de/fusion/members/friederike-klan/
§ www.fusion.cs.uni-jena.de/fusion/members/birgitta-konig-ries/

DOI: 10.2312/eurovisshort.20181081

conferences organized by German and international biodiversity organizations (check supplementary material). The online survey was active from August 2015 until December 2017. Its preview is available at https://tinyurl.com/yb2ysyuu. Besides, a commentary paper [KGKR16] along with a survey link was also published in an international journal to reach a large audience. We have received 100 responses in total. Considering the outreach of participants through all these venues this number is low. This is symptomatic for the limited willingness to share knowledge across interdisciplinary borders. Within the survey, some questions were multiple choice and others were single choice. For many questions a commentary section was provided to allow the participants to provide additional information and viewpoints on different inquiries. For the convenience of the participants in completing the survey, no mandatory fields were added. This resulted in many questions remaining unanswered. Therefore, the scores calculated and presented in Section 3 are based on the number of answers for each question received rather than the total number of survey responses.

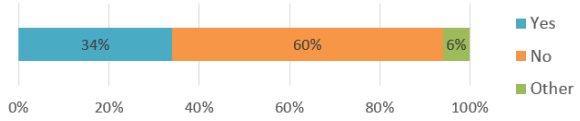


Figure 1: *Do users find it difficult to select a visualization for presenting their research data? The total number of responses received were 100.*

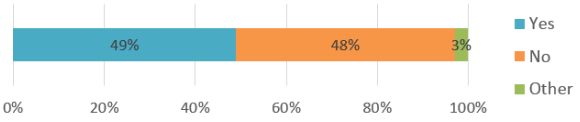


Figure 2: *Are users interested in having a software tool that can guide them in the selection of suitable visualizations? The total number of responses received were100.*

3. Results and Discussion

3.1. Issues with visualization selection

Figure 1 shows that the majority of biodiversity researchers feel comfortable with their visualization skills and indicate to not face problems when selecting and creating visualizations. On the other side, the study participants have expressed (Figure 2) the need for a visualization support tool that can assist them in these processes by recommending suitable visualizations. Through comments, they have directed their concerns on various issues they face when choosing a proper visualization. In the following, we have analyzed these comments and have categorized them into distinct visualization selection challenges:

- **Visualization selection dilemma:** The participants face difficulties to find the best visualization solution to represent their data. Nowadays with ample of visualizations available, an appropriate visualization selection can become challenging as for a visualization layman, every other visualization looks the same.
- **Dependency on the visualization publication medium:** The participants find it more complicated to publish visualizations in journal articles, as it is costly to use colors. Whereas for online presentations, users have a wide selection and choice of visualizations which they can easily configure to make them more appealing to their audience.
- **Lack of knowledge:** The participants feel that they are unaware of alternative types of visualization techniques. Their visualization selection options are limited to what they have developed earlier or what they have seen in previously published work. Due to this, they use similar visualization types repetitively.
- **Visualizing large and complex datasets:** The participants find it difficult to choose suitable visualizations to represent large and complex datasets. It is problematic to convey a message within multi-dimensional datasets clearly and precisely using a single figure.

3.2. Visualizations and their usage in the biodiversity domain

To this end, participants were shown a list of different visualizations and were asked to indicate the different purposes for which they use these visualizations in their daily work. This list was produced after knowing the types of common visualizations available in the biodiversity publications. In order to get a varied result, participants were asked to provide the answer to this question in a form of free-text. Table 1 shows the most frequently used visualizations and its usage. The word cloud associated with each visualization shows the usage or purposes indicated by the study participants. The larger the size of the word is, the more frequently it was mentioned by the participants. It is evident that biodiversity scholars use a spectrum of different visualizations for similar tasks, for example, the representation of data grouping and its comparison is done by scatterplot, boxplot and bar chart. However, there are typically only one or two tasks that are prominent to each visualization. Scatterplot for example, is used to illustrate the result of a principal component analysis (PCA) or to visualize the spatial distribution of objects, e.g. species. Dendrograms are frequently used for facilitating phylogenetic or a cluster analysis. In the supplementary material, we have provided the complete list of considered visualizations and their uses. The study participants were also asked to provide the reason for not using some of the visualizations listed. We have categorized these reasons into two groups : Never Needed and Don't Know (not aware of the visualization). Figure 3 indicates that Parallel Coordinates, Treemap, Venn Diagram and Coplot (conditioning scatterplot) are much less used compared to the other visualizations, although at least half of the respondents were aware of those types of visualizations. This raises question why those visualizations were not considered although most of them are more advanced and suited to multidimensional data. As it turned out, participants consider parallel coordinates as difficult to interpret and hard to comprehend. One participant said that instead of it he will prefer to represent different dimensions via different 3d plots. The study participants also noted that one of the reasons for rarely using treemaps is because often it is dynamic and is thus hard to include it in a paper. Some participants show more preference to lattice graphs rather than coplot. The reason for the rare use of Venn diagrams is that they are mostly known to represent concepts or ideas rather than numerical data. One participant said that he would like it when its area and colors were also meaningful.

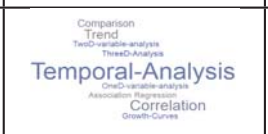
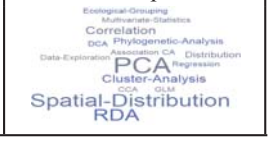
BarChart	PieChart	LineChart	Heatmap	Coplot
				
ScatterplotMatrix	Boxplot	DensityPlot	Dendrogram	Scatterplot
				

Table 1: *Visualization types and the purposes they are used for in the biodiversity domain*

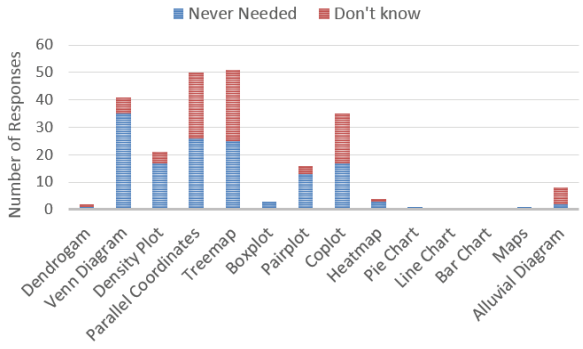


Figure 3: *Stacked column chart showing the number of participants that never used a given visualization or were not aware of it yet.*

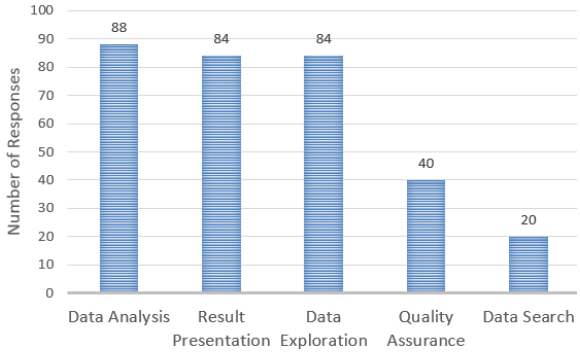


Figure 4: *For what purposes do you use visualization? Total responses received 99. Multiple answers per participant possible.*

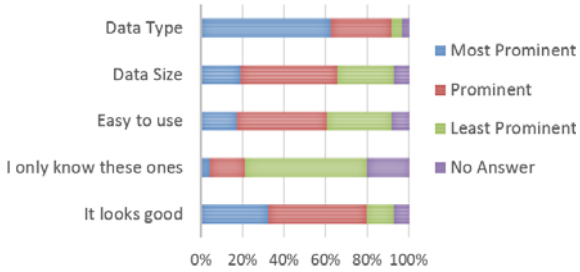


Figure 5: *What factors does users consider while selecting a particular visualization for their task? Total responses received 99. Multiple answers per participant possible.*

3.3. Visualization tool requirements

In the following, we have analyzed and have categorized the comments from the participants about their expectations on a software tool that supports visualization:

- **Visualization support for data management tasks:** We found out that visualization usage is not limited to the purpose of presenting results. There is a high necessity of visualization support for other data management related tasks. Our results (Figure 4) reveal that data analysis, result presentation and data exploration are the three prominent tasks that can be effectively supported by visualization. Moreover, Figure 4 also conveys that researchers have started realizing the usefulness of visualization for data quality assurance and data search.
- **Factors for visualization selection:** Visualization tools offer visualization selection based on certain factors. In our study (Figure 5) we found that scientists consider these three factors as most prominent for visualization selection: data type, aesthetics and data size. Here aesthetics refers to the clarity and comprehensibility of a visualization. Comparing these factors to the ones presented in [RDB14] they are quite similar with an addition to the factor 'Ease of use' which our participants have indicated fairly equally important as 'Data size'.
- **User centric:** The participants feel that a visualization software tool should not be too prescriptive or conditional. It means that it should provide a range of options (solutions) to the users to select from instead of selecting one for them. This also means that the solutions (or visualization recommendations) should not be fixed to and based on some preset conditions within the software. The software should be adaptive to integrate user responses or preferences in a real-time and then provide a personalized solution.
- **Easy to use:** The participants expressed their needs for a visualization software that is easy to access, to use and to understand. Instead of making a user guess on what procedure to follow in order to create a visualization, the software should guide the user at each step.

- **Showcasing:** The participants believe that showcasing what visualizations are present in the system will make them aware of the different options available within the tool. Such a showcase can be implemented in the form of a visualization knowledge-base website or a guidebook. This can assist them in efficiently exploring, interpreting and developing graphical representations of their data.
- **Interactivity:** The participants consider interactivity as an important feature of a visualization. A visualization software tool should offer support for interactive visualizations. Interaction within a visualization helps to explore the different data dimensions, gives a better overview of a visualization and its elements, provides visualization customization and enables the audience to engage with the visualization.
- **Multi-platform support:** The participants indicated that the visualizations produced by software tools should be flexible and platform independent. This means that visualizations should be easy to export or import and should not depend on any graphical tool. They can be easily altered by other graphical platforms or tools.
- **Color-blinded friendly:** The participants want visualization tools to produce color-blind friendly visualizations so that color-blinded community can effectively use them too. A color-blind person has trouble seeing red, green, blue or mixtures of these colors. So, the visualizations produced for them either avoid such color combinations, includes both textures and patterns instead of only colors, uses colors with high contrast, leverages symbols wherever possible or make use of special color-blind friendly color palettes.
- **Visualization audience:** The participants also consider the audience of the visualization as one of the important factors in the visualization selection process that needs to be considered within the tool. The selection of a visualization will be different if the visualization is going to be presented to graduate students, experienced scientists, layman or stakeholders etc.

3.4. Data exploration workflow

One of the important data management tasks for which visualization is used is data exploration (Figure 4). Data exploration provides a sneak peek into the data at hand and thus helps in making an initial decision about the relevance of a dataset for answering a certain research question. What steps need to be followed to get an initial exploration and understanding of the data? We have asked our survey participants to provide their experience about how they explore a dataset. We have summarized their answers into the following four major steps:

- **Data Investigation:** In this step, if the dataset is not clean then one would investigate data for various quality issues like missing data, data inconsistencies etc. and perform necessary cleaning. Then one would further examine the different features of the data (data dimensions, data size, data types etc.).
- **Data Overview:** In this step, one might perform the following actions: getting an overview of the dataset via different multi-dimensional visualizations, examining the distribution of the data to understand if it is skewed or symmetric, detecting outliers, summarizing the data for further statistical analysis or refinements.
- **Data Refinement:** In this step, one might perform the following actions: filter or subset the data based on the individual analysis goal, transform the data (for example at different scales to remove skewness), create the derived or compound variables as per the analysis requirements, remove outliers if those were spotted in the previous step.
- **Data Analysis:** In this step, one might perform the actual analysis tasks like hypothesis formulation, understanding relationships existing within a dataset, doing comparisons etc.

These are the preliminary steps that researchers follow to do an initial exploration of the data wherein different visualizations are needed to facilitate each step. Data investigation is the foremost step that users perform. Then depending on the individual goals, some of the remaining steps follow in non-particular order. For example, if a user has some information about the data then the user will go for data refinement to explore the variable of its interest. Whereas, if a user has no prior information about the data, then the user might be interested in viewing a multi-dimensional view to get an overview of the complete dataset and then can choose the variables of interest. After the refinement step, the user might be interested in getting an overview or in summarizing the variables of interest and then would want to do further analysis. Again, after analysis, the user might perform further data refinement or might be interested to get an overview of the altered dataset as per its individual requirements.

3.5. Conclusion

Our study revealed that although biodiversity researchers feel comfortable with their current visualization practices, they wish to have a software support in order to choose proper visualizations to represent their data. Major challenges arise from the large number of visualizations available today and from the increased size and complexity of the data to visualize. We have observed that apart from using visualization for data presentation and analysis, users now realize the usefulness of visualization for other data management tasks like data exploration, data search and quality assurance. Thus opening up a research dimension for the visualization community to provide visualization as a service to the data management process at its different stages. The requirements for such a visualization support tool include the possibility to showcase available visualizations, interactivity and multi-platform support. These will be considered in our ongoing research on the construction of a visualization recommendation tool for the biodiversity community.

4. Acknowledgment

The work has been funded by the DFG Priority Program 1374 "Infrastructure-Biodiversity-Exploratories" (KO 2209 / 12-2).

References

[CSC02] Cooper R. J., Schriger D. L., Close R. J.: Graphical literacy: the quality of graphs in a large-circulation journal. *Annals of emergency medicine 40*, 3 (2002), 317–322. 1

[DLW*17] Dasgupta A., Lee J.-Y., Wilson R., Lafrance R. A.,

CRAMER N., COOK K., PAYNE S.: Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 271–280. 1

[KGKR16] KAUR P., GAIKWAD J., KÖNIG-RIES B.: Towards recommending visualizations for biodiversity data. *Biodiversity and conservation 25*, 9 (2016), 1801–1803. 2

[KKUW07] KULYK O., KOSARA R., URQUIZA J., WASSINK I.: Human-centered aspects. *Human-centered visualization environments* (2007), 13–75. 1

[RDB14] ROUGIER N. P., DROETTBOOM M., BOURNE P. E.: Ten simple rules for better figures. *PLoS computational biology 10*, 9 (2014), e1003833. 3

[SSS*06] SCHRIGER D. L., SINHA R., SCHROTER S., LIU P. Y., ALTMAN D. G.: From submission to publication: a retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the british medical journal. *Annals of emergency medicine 48*, 6 (2006), 750–756. 1

[WMWG15] WEISSGERBER T. L., MILIC N. M., WINHAM S. J., GAROVIC V. D.: Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS biology 13*, 4 (2015), e1002128. 1