Experiences in the Development of a Measure and Indicator Web-Based Visualization System

Alex Baumann, Mary Beth Smrtic, Andy Dufilie, William Mass, Georges Grinstein

University of Massachusetts Lowell, Lowell MA

ABSTRACT

We have designed several high-end desktop visualizations for the government and industry, including our Universal Visualization Platform (UVP) currently being used in high-dimensional data analysis [1-4]. With the formation of the Open Indicators Consortium [5] we are developing a similar system incorporating much of what we have learned, but as an open-source high performance web-based system specifically designed to analyze and display measures and indicators that are collected by the consortium members. Members and their users found the visualization tools to be very useful in pointing out anomalies, patterns and errors in their indicator data. We briefly discuss the development process and a few of these discoveries.

KEYWORDS: Web-based Visualization, Measures, Indicators, Open Source, Data Errors, Anomalies, Agile Development

1 INTRODUCTION

The Open Indicators Consortium (OIC) was established in the fall of 2008 to create user friendly, open-source, high-performance web-based software tools for data visualization to serve regional planners, non-profits and local and state agencies. These organizations have distinct missions that overlap in their needs for data management, analysis and visualization. Each Consortium Member (CM) has a constituency of users that range from novice users (typically the general public) to experts (public planners).

The University of Massachusetts Lowell leads the OIC software development effort and the CMs are metropolitan, regional or state partnerships across the country, including Atlanta, Boston, Chicago, Columbus, Arizona, Connecticut and Rhode Island. All CMs can be described as *intermediary organizations* in that they serve both the public and other constituent groups. As intermediary organizations the CMs identify relevant measures and indicators, and the extent of available metadata, the data analysis and desired flexibility in visualizations of their data appropriate to a range of their users from novice to advance analyst. As a mutual learning opportunity, the OIC facilitates discussion of better or best practices, including approaches and solutions to common or related problems of data selection, analysis, presentation formats and visualizations.

2 INTERACTING WITH OUR CONSORTIUM MEMBERS

Although most CMs are familiar with GIS, developing a software system is new to them and their experience with software developers and highly interactive tools is quite limited. We selected the Agile iterative development process in order to provide them with usable systems as frequently as possible. This allows CMs to better understand the features of the software and to help ensure that their requirements are met. Discoveries made by the members and their constituency helps to drive the development in terms of bug fixing and feature requirements. It is critical to have this group participate in the development, as our team at UMass Lowell, are not lead indicator users. We also do not know their user base, which will be the eventual target audience of the system. There are in fact multiple users: the developers of the CM's site; the analysts, statisticians and planners who are fairly knowledgeable about GIS systems; the trainers who develop tutorials and tools to use the CM's web site; and finally the end user, the public, who typically will range from expert computer users to true novices.

3 DATA ERRORS

Exploratory visualization is useful in learning about a dataset. Our consortium member's datasets have differing levels of accuracy and completeness. Our members been surprised at how useful our tools have been in pointing out data errors that may otherwise have gone unnoticed.

Figure 1 shows counties in Connecticut colored by latitude. This simple visualization highlighted errors in the data to the group from Connecticut. Even if you know very little about latitude, you can see that the teal colored county in the middle is out of place. The Columbus CM had a similar data issue in which certain data points were mapped to census "county subdivisions", while others were mapped to census "places", which was a problem they were not aware of. These errors turned out to be easily fixable, but most likely would not have been found if not for the visualization and software tools.

The left half of Figure 2 shows US counties colored by an indicator labeled *Percent Cuban*. One member pointed out that Miami should have one of the highest percentages of Cubans, so we compared the numbers to those found on the Miami-Dade county website. We found that the field labels were incorrect in the database. The right side of Figure 2 shows the corrected data. If you look carefully you can see a dark red county in southern Florida. That is Miami-Dade County and has the highest Percent Cuban in the United States at 28.8 percent. As the development focus was on technology more so than data, such errors may not have been realized without user feedback from users with some domain knowledge.

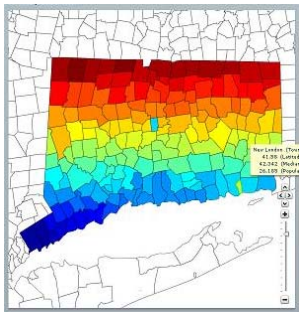

Figure 1. Connecticut counties colored by latitude

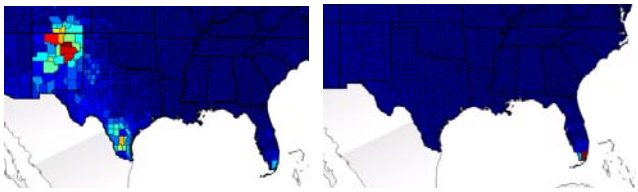

Figure 2. United States counties colored by a field incorrectly labelled as Percent Cuban (blue: low percent red: high percent)

4 Outliers

When creating a new visualization it is common to first check outliers to see if they are reasonable. Figure 3 shows the United States census data at the county level, with one clear outlier in the top right of the scatter plot. Without knowledge about the data itself, it can be difficult to determine whether or not this outlier is valid or not. One of the CMs looked up this outlier (Kalawao County, Hawaii) on the Internet found that it was, until 1969, a community of persons suffering from leprosy. It is separated from the rest of the island by cliffs over a quarter mile high and was under quarantine until 1969. At that time the government lifted the quarantine but would not allow new residents into the colony [6]. This explained the high median age and clarified its validity as an outlier in the dataset. This discovery process led to an added requirement, namely the ability to link to a general search engine based on indicator metadata and values.

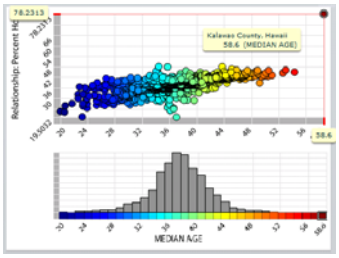

Figure 3. United States Counties Median Age Distribution

5 Surprises in the Data

As we were developing the ability to show multiple layers on a choropleth map, a high priority requirement for the Boston group, we provided them with a map of foreclosures point data in the Boston area layered over census tract shapes that could be colored by an indicator. Figure 4 shows the census tracts colored by *Races: Percent Black* with white dots where there were foreclosures. Boston had reason to expect a disproportion of foreclosures in low income and minority communities, but did not expect such a stark correlation. The map shown in Figure 5 reflects a markedly high concentration of foreclosures in African-American communities, even when compared to other minority groups such as Latin Americans. This proved the utility of our tool for determining if some communities are particularly impacted by a crisis, whether a community is defined geographically, by ethnicity, by income or by some combination of indicators and measures.

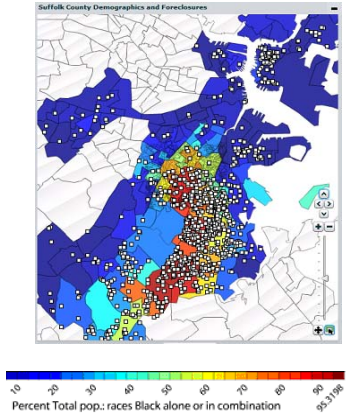

Figure 4. Foreclosures (white dots) and Percent Black population (colored in census tracts) in the Boston area

6 Impact of Visualization Tools

Besides identification of errors in the data or checking of data for validity, many CMs have been excited about the ability to compare across different regions for use in benchmarking. The Columbus CM used the US County dataset to explore similar counties to their region (Franklin County, Ohio). Counties with a similar median age were selected and other indicators were explored on a scatter plot. This allowed exploration of how these counties compared to others across the country, as shown in Figure 5. This selection of counties was also used as a subset of the data to identify how these similar counties differed from one another in terms of other variables.

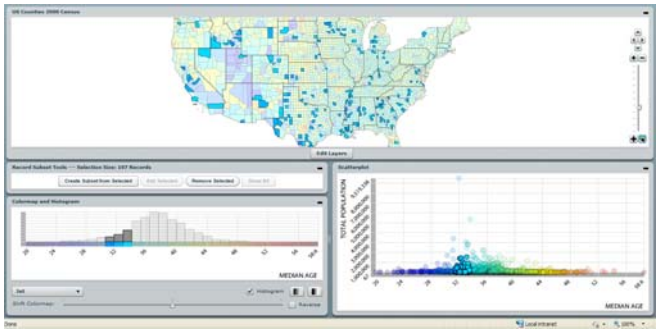

Figure 5. US County data map showing counties deemed similar to Franklin County, Ohio by the Columbus CM.

7 Conclusion

Data visualization was new to many of the CMs and most had not seen their own data visualized before. Some were familiar with Excel, or had websites with static choropleth maps, but were excited to be able to explore their own data dynamically. The new tools allowed the CMs to gain two main insights; first that their data was often neither clean nor complete, and second, that additional data or more complete data would provide much more impact.

References

[1] Gee A., H. Li, M. Yu, M.B. Smrtic, U. Cvek, H. Goodell, V. Gupta, C. Lawrence, J. Zhou, C.-H. Chiang, G. Grinstein (2005). Universal Visualization Platform. In Robert F. Erbacher, Philip C. Chen, Jonathan C. Roberts, Matti T. Gröhn, Katy Börner (Eds), Visualization and Data Analysis 2005 (San Jose, California, January 17-18), Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 5669, pp. 274 - 283.

[2] Sharko, J. and G. Grinstein, Visualization Fuzzy Clusters Using RadViz, to appear in the proceedings of the Information Visualization in IEEE Biomedical Informatics Symposium held in conjunction with the 13th International Conference on Information Visualization, Barcelona, 14 - 17 July 2009.

[3] Konecni, S., G. Grinstein, and J. Zhou, A Visual Analytics Model Applied to Lead Generation Library Design in Drug Discovery, to appear in the proceedings of the IEEE Visual Analytics Symposium held in conjunction with the 13th International Conference on Information Visualization, Barcelona, 14 - 17 July 2009

[4] Goodell H., C-H. Chiang, C. Kelleher., A. Baumann , and G. Grinstein (2006), Collecting and Harnessing Rich Session Histories, International Conference on Information Visualization (IV06), London, July 5-7, 2006

[5] http://www.openindicators.org

[6] http://en.wikipedia.org/wiki/Kalawao_County,_Hawaii