An Interactive Mapping System Incorporating Data Reliability Information

Min Sun  
George Mason University  
4400 University Dr, Fairfax, VA, 22030  
msun@gmu.edu

David Wong  
George Mason University  
4400 University Dr, Fairfax, VA, 22030  
dwong2@gmu.edu

Barry Kronenfeld  
Eastern Illinois University  
600 Lincoln Ave, Charleston, IL  
bjkronenfeld@eiu.edu

ABSTRACT
Data quality should be considered in compiling maps in order to reveal reliable information about the spatial variation of a phenomenon. However, creating classes in a choropleth map by maximizing data reliability (i.e. the statistical differences of observed values between classes) often lead to useless maps with very uneven number of observations. An interactive mapping system is developed to allow users to consider more than just data quality in determining class breaks to create maps which are relatively reliable and more effective in spatial pattern detection.

Keywords  
Interactive mapping, visual analytics, data quality, class separability

1. MAPPING WITH ERROR
Choropleth (classified) maps are often designed to explore spatial patterns, but determining value classes (i.e. class breaks) is critical in revealing the information captured in data. Compiling a truthful map is in no way simple even when data are accurate because multiple factors are involved and interact in the process. The presence of error makes such process more challenging. Sun and Wong et al. [1] pointed out that as values being maps are often statistical estimates which have a certain level of uncertainty (as reflected by standard error), these estimates assigned to different classes in a choropleth map are not different statistically. Thus, maps may show spatial patterns, which are the results of systematic differences in values over space, but in fact the patterns may not exit if values in different classes are not really different [2]. Unfortunately, most spatial data are gathered from surveys, and therefore the estimates have uncertainty. Such uncertainty information should be reported in maps and incorporated in the map compilation process to create more reliable map [3].

When information about spatial data quality is available, most studies focus on using cartographic symbols and statistical plots to represent the reliability information of mapped values, e.g. [4-7]. However; these methods merely display error information, warning readers about the reliability of the maps, but cannot offer more reliable maps. The quality information is not used in designing how the mapped values should be presented. Recently, Sun and Wong et al. [1] proposed a class separability measure to evaluate how estimates are statistically different across classes. Each pair of estimates ($X_i$ and $X_j$) on two sides of a potential break has a confidence level ($CL$) which reflects the likelihood that the two estimates are statistically different. In Eq. 1, $SE$ is the standard error of estimate $i$ and the function $\Phi$ returns the probability of a z-score of the estimate difference. The minimum $CL$ of all pairs of estimates between two classes $A$ and $B$ is defined as the separability measure associated with the respective class break (Eq.2).

$$CL_{i,j} = \Phi \left( \frac{|x_i - x_j|}{SE_i^2 + SE_j^2} \right)$$

$$S_{A,B} = \min_{i \neq j} (CL_{i,j})$$

Based on the class separability measure, Sun and Wong et al. [1] proposed the “class separability” method to determine class break values by choosing the break points with the highest $S_{A,B}$ values. The resultant map will have the most “separable” classes statistically. Fewer classes will have higher separability levels among them, but more classes will have lower separability levels. Empirically, break points with higher separability levels tend to be at the extremes of the distribution. Accordingly, resultant classifications are often unbalance with many values in middle class(es) but very few estimates fall into class(es) at both ends. The resultant map may not be too useful to reveal spatial patterns as the majority of the observations are assigned into one or a few classes. Thus, a compromise is to adjust results from using the class separability methods, while maintaining the high separability levels between classes to the largest extent. One potential compromise is to divide the “large” class(es) into smaller but more balanced ones so that the underlying spatial pattern in the data can be revealed more effectively.

This paper and the demo$^1$ describe our effort to develop an interactive mapping system with new graphs and user interaction design. The system enables map makers to explore how the class break values can be adjusted to maximize the statistical differences between classes on the one hand, and to maintain an acceptable level of separability among different classes. The ultimate objective is to create relatively reliable maps that can reveal spatial informative effectively.

2. INTERACTIVE MAPPING SYSTEM
To produce maps with more balanced classes, maximizing separability should not be the only criterion to consider. Other factors besides data reliability are important. However, the data reliability criterion sometimes conflicts with other factors, especially the evenness of observations across different classes, an important factor for creating maps that can reveal spatial clusters. One way to resolve the conflicts between mapping criteria is to let make makers to decide the trade-off among criteria. In other words, the separability levels of the map has to be lowered, implying a less reliable map, to accommodate other criteria. But how much reliability should to give up in order to attain higher levels in other criteria? Such trade-off is often unknown to the map maker until different levels of the criteria are explored. Therefore, the process of exploring different criterion values is

$^1$ https://youtu.be/2WMBaTVIUyk
actually a process of evaluating those trade-offs through a series of experiments [8]. Repeating the experiments is best in an interactive visual-based system, including the following capabilities:

- Data visualization: different forms of representation with each concentrates on one or multiple aspects of information (including data reliability) that needs to be considered in the process in determining class break values.
- Linkage: multiple forms of data representation need to be linked together so that information updated in one form is reflected in the content in another form. Underlying correlation may be revealed through the synchronized changes.
- Interaction techniques: an important element of a heuristic mapping process is to capture user’s inputs. This capability enables the user to repeat the experiment and to determine the most “desirable” solution according to the particular purpose of the application.
- Real-time computation: On-the-fly statistical computation is required to be in the same environment where data are presented.

Figure 1 shows the four major components in our interactive mapping system. A map window (Figure 1 a) displays the intermediate and resultant maps that are created during the heuristic experimental process while the other windows are used to drive mapping process with data reliability in consideration.

The bar plot (Figure 1 b) is to show the distribution of estimates and associated errors. This provides the environment to implement the class separability concept to maximize the statistical differences between estimates in different classes (Sun and Wong et al. 2014). By dragging the slider bar underneath the bar plot, users can adjust or determine the minimal class separability level acceptable. All break values with separability levels higher than the acceptable level will be selected. Because this method considers only the separability measure, the classification is likely very unbalanced.

To reduce the unbalance of distribution of estimates across classes, we allow users to manipulate break values (Figure 1 d). If a class has too many estimates, the user can insert a new break value at an “appropriate” position after evaluating the distribution of estimates shown by the bar plot. If two break values are too close such that the class in-between has too few observations, one of the break values with lower separability level may be removed. Where break values should be added and which break points should be removed are based on the distribution of estimates. Users can also move the break value to the left or right to make slight adjustments after a break point is placed. Maps will be automatically updated to display the results after an adjustment of the class break value which leads to a new displayed the spatial pattern and reliability of the classification.

Besides manipulating break points in an existing classification, users may consider more than one mapping criteria to create more balanced and effective classes to detect spatial patterns. A star plot (Figure 1 c) is used to reveal the underlying relationships among different mapping criteria. All feasible classification schemes are identified and represented by polygons in the star plot. Each scheme is equivalent to the combination of values referring to different mapping criteria. If equal weights are given to all criteria, user can select the combination with moderate values for all criteria. By clicking that corresponding polygon, a new map will be generated based on that chosen scheme. The user can explore and compare all possibilities and determined the most desirable one by selecting all options one after the other. In addition, the user can always manipulate the break values to adjust classes.

The details about the components and symbology in the graphs in the system and the user interaction enabled by the system are described in the attached demo1.

3. EXPERIMENT
To demonstrate the mapping process enabled by the system, we used a dataset downloaded from SEER Stat2 database which include the error information (SE). The mortality rates of whites for all causes of death (COD) for the 768 health service areas in the continental U.S. from 1969 to 2011 are used in the demo.

4. CONCLUSION
Maps offer decision support by identifying spatial patterns, including hot-spots of specific health outcomes, events or crime incidences. However, without considering error in data, the identified patterns or hot-spots can be erroneous. While the separability concept helps determine highly separable classes, the proposed mapping system can improve the class balance of maps originally compiled using the class separability criterion. These improved maps are more informative about the spatial patterns although the separability levels between classes have to be compromised.

5. ACKNOWLEDGMENTS
Research reported in this publication was partly supported by the National Institutes of Health (NIH) under Award Number R01HD076020. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

6. REFERENCES

1 https://youtu.be/2WMBaTVIUyk
2 http://seer.cancer.gov/seerstat/


