A comparison of optimal map classification methods incorporating uncertainty information

Yongwan Chun*[1], Hyeongmo Koo[1], Daniel A. Griffith[1]

[1] University of Texas at Dallas, USA

*Corresponding author: ywchun@utdallas.edu

Abstract
Uncertainty in spatial data attributes can produce unreliable spatial patterns in choropleth maps, but only a few studies have considered uncertainty in map classification processes. Unfortunately, a less desirable classification result often is generated by existing methods. For example, most observations are assigned to a single class while the remaining classes have a very small number of observations allocated to them. Also, selection of proper criteria for an optimal map classification is difficult. The purpose of this paper is to expand the discussion about incorporating data uncertainty for map classification by extending optimal map classification strategies with Bhattacharyya distance. The proposed method is illustrated with an application of soil lead contamination measurements in the City of Syracuse.

Keywords
Uncertainty, Map classification, Choropleth map

I INTRODUCTION

Most spatial data attributes inevitably contain uncertainty due to sampling and/or measurement error, among other sources of error (e.g., specification). The preferred form of visualization for these data is uncertainty as well as attribute values, which is limited because of existing technical and conceptual deficiencies. One common approach is to utilize a bivariate mapping technique that simultaneously represents estimates with their corresponding uncertainty information using additional visual variables. For example, using Bertin's (1983) graphic variables, such as color value and texture, with a general choropleth map (e.g., MacEachren et al., 1998, 2005; Xiao et al., 2007).

Sun et al. (2014) propose a class separability measure to incorporate uncertainty information in a map classification to produce more reliable spatial patterns. This approach produces a more elaborated classification result. But it is heavily affected by outliers, and, at worst, results in a number of classes with a single observation, allocating most observations to a single class. Sun et al. (2016) further propose a heuristic approach to overcome this imbalance issue of the class separability classification results; but untrained users still might find selecting values of criteria for achieving an optimal map classification to be difficult.

The purpose of this paper is to expand the discussion about incorporating data uncertainty for choropleth mapping. More specifically, this study incorporates optimal map classification strategies with Bhattacharyya distance.

II METHOD

Bhattacharyya distance is effective to quantify dissimilarities between two probability distributions (Bhalerao and Rajpoot, 2003), resulting in it commonly being used in feature selection and extraction (e.g., Choi and Lee, 2003; Reyes-Aldasoro and Bhalerao, 2006). When two observations conform to a normal distribution, Bhattacharyya distance between the two observations can be calculated with the following equation (Coleman and Andrews, 1979):

$$D_B(i,\ j) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{SE_i^2}{SE_j^2}+\frac{SE_j^2}{SE_i^2}+2\right)\right)+\frac{1}{4}\left(\frac{\left(\bar{x}_i-\bar{x}_j\right)^2}{SE_i^2+SE_j^2}\right) \quad (1)$$

where $\bar{x}_i$ and $\bar{x}_j$ are the estimates and SE_i and SE_j are corresponding standard errors for observations i and j. Bhattacharyya distance can index the dissimilarity between two observations having the same estimate with the first term of equation (1). In general, an optimal classification can be achieved by minimizing within class Bhattacharyya distance. A within class distance may be determined in two different ways. The first is the sum of pairwise distances within a class (SB), and the other is the maximum pairwise distance within a class (MB). Although the first method calculates a measure from all pairwise distances, the second method intends to control for the worst case to improve homogeneity within a class. Class breaks can be determined in order to minimize the sum of costs, which are within class distances here.

This study utilizes optimal classification methods incorporating uncertainty information by extending the optimal classification method developed by Cromley (1996). He shows that an optimal univariate map classification can be modelled as an analogy to a constrained shortest path problem, when an acyclic network is constructed by the rank ordered estimates of observations. In this network, a node represents an observation, and lines correspond to assigning observations to a class. Here, a cost (or impedance) of a line is defined by within class Bhattacharya distance. An optimal classification can be constructed by identifying a combination of lines minimizing a total cost.

Extending Cromley (1996), optimal map classification can be formulated as follows:

Minimize:	$\sum_{i,\,j} c_{i,j} d_{i,\,j}$	$\forall\ i,\ j\ \in N, i \le j$	(2)
Subjected to:	$\sum_i d_{i,\,k} = \sum_j d_{k,\,j}$	$\forall i\ \in In_k;\ j \in Out_k$	(3)
	$\sum_j d_{s,\,j} = 1$	$\forall j \in Out_s$	(4)
	$\sum_i d_{i,e} = 1$	$\forall i \in In_e$	(5)
	$\sum_{i,\,j} d_{i,j} = T$	$\forall\ i,\ j\ \in N, i \le j$	(6)
	$d_{i,\,j} \in \{0,1\}$	$\forall\ i,\ j\ \in N, i \le j$	(7)

where $d_{i,\,j}$ is a binary decision variable, $c_{i,j}$ is the cost of a line from i to j, In_k is a set of lines that terminate at node k, Out_k is a set of all lines that originate at node k, and nodes s and e respectively are the minimum and maximum values. Equation (2) is the objective function, and equation (3) ensures that all observations are assigned to a class. Equations (4) and (5) ensure that the first node is assigned to the first class, and the last node to the last class. Equation (6) constrains the number of classes, and equation (7) is a binary integer restriction.

III APPLICATION

The proposed method is illustrated with lead (Pb) contamination measurements collected from soil samples across the City of Syracuse (Griffith, 2008). These measurements are in milligrams per kilogram of soil (ppm), and are log-transformed so that their frequency distribution better conforms to a bell-shaped curve. A Pb surface was krigged with a Bessel function semivariogram model (26,402 pixel values were interpolated), and then the krigged surface was aggregated to 264 grid cell polygon values. Predicted measurements and prediction standard errors for these grid cells are utilized for map classification purposes. The proposed methods have been implemented as an extension of ArcGIS 10.1 using C# in the Microsoft .Net Framework 4. The optimization problems are solved by a branch-and-bound algorithm using Gurobi optimizer 6.5.0.

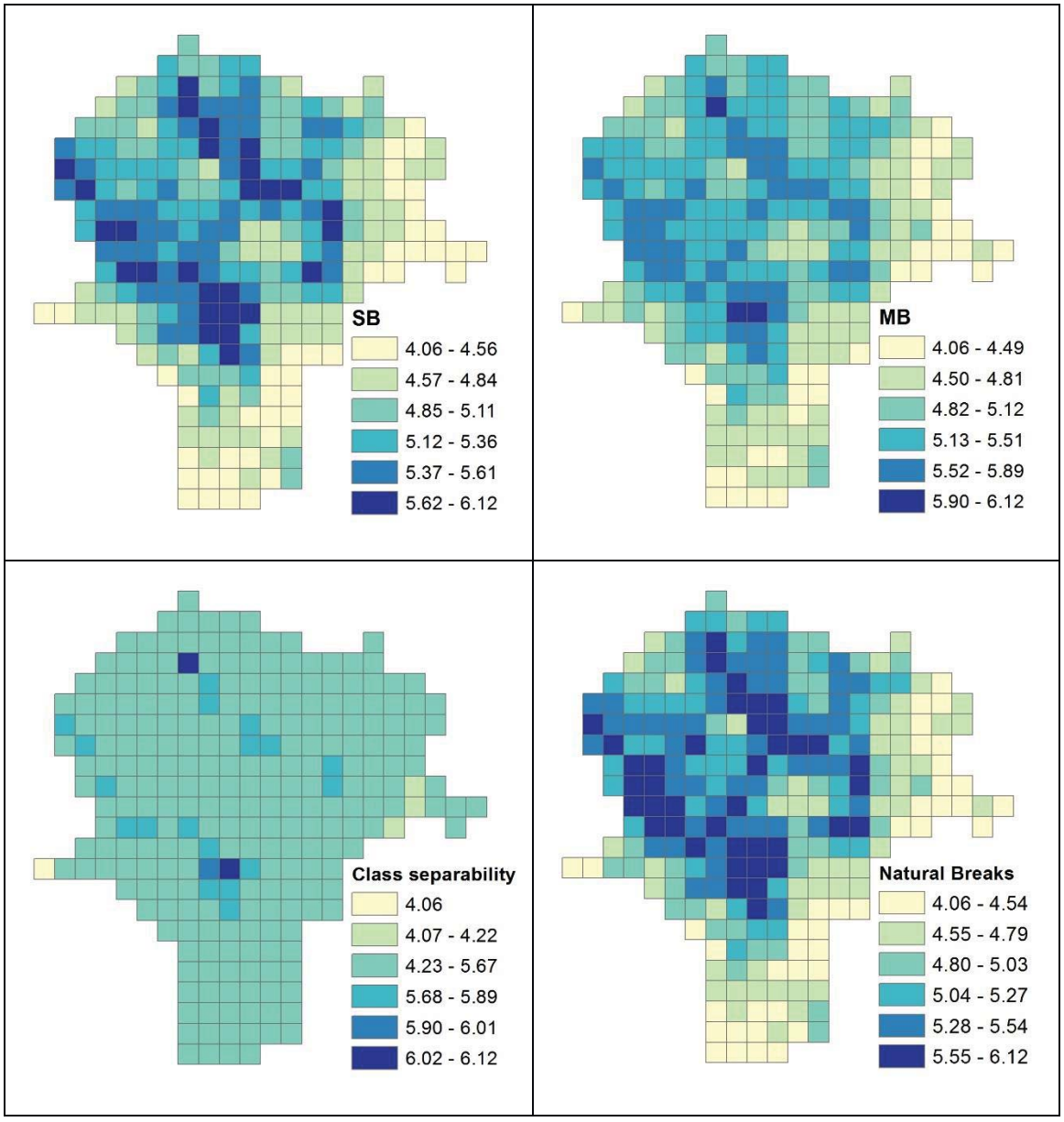

Figure 1: Map classification results for an aggregated Pb krigged surface across Syracuse

The classification results are compared with those from the class separability method by Sun et al. (2014), and from the standard natural breaks classification method. Figure 1 displays map

classification results for Pb levels. The class separability method produces a spatial pattern that is remarkably different from the other results. This method produces classes that contain only outliers on both sides of its frequency distribution. The other three map classifications show a relatively similar spatial pattern, although they have a noticeable difference for the last class. The last class by the natural breaks method has the greatest number of observations among the map classification results. In contrast, class 6 in MB has only three observations.

Table 1 presents the numbers of observations for output classes. The results of the class separability method display considerable variation in terms of the number of observation counts across classes. More specifically, most of the observations are assigned to class 3. However, classes 1 and 5 have only a single observation. In contrast, proposed classification methods have a more balanced number of observations across classes. The natural breaks classification also produces well balanced results, but this method considers only variances among estimates without a consideration of uncertainty (Jenks, 1977).

Table 1. The number of observations in classification result classes

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
SB	43	56	46	48	44	27
MB	32	59	55	76	39	3
Class separability	1	3	239	18	1	2
Natural breaks	38	50	42	43	50	41

IV DISCUSSION

This study extends map classification by integrating an optimal map classification (Cromley, 1996) and uncertainty information with various costs (e.g., distance here) and objective functions. Generally, the proposed methods produce homogenous classes by their respective criteria, and also achieve visually balanced classification results. Some limitations should be investigated in future studies. First, a performance evaluation for the proposed methods needs to be investigated. Because widely used evaluation methods examine only estimates (e.g. Jenks and Caspall, 1971; Xiao et al., 2007), these methods are not effective for evaluating the performance of the proposed map classification methods. Second, the proposed map classification methods are solved using a branch-and-bound algorithm. Thus, as the number of nodes increases, the amount of time needed to find an optimal solution tends to increase often at roughly an exponential rate.

V ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health, grant 1R01HD076020-01A1; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Institutes of Health.

References

Bertin, J. (1983). *Semiology Of Graphics: Diagrams, Networks, Maps*. Madison: University of Wisconsin press.

Bhalerao, A. H., Rajpoot, N. M. (2003). Discriminant feature selection for texture classification. *In Procedings Of The British Machine Vision Conference 2003*. United Kingdom.

Choi, E., Lee, C. (2003). Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, 36(8), 1703–1709.

Coleman, G. B., Andrews, H. C. (1979). Image Segmentation by clustering. *In Proceedings of IEEE,* 67, 773–788.

Cromley, R. G. (1996). A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographical Information Systems*, 10(4), 405–424.

Griffith, D. A. (2008). Geographic sampling of urban soils for contaminant mapping: how many samples and from where. *Environmental Geochemistry and Health*, 30, 495-509.

Jenks, G. F. (1977). *Optimal Data Classification for Choropleth Maps*. Occasional Paper No. 2, Department of Geography, University of Kansas.

Jenks, G. F., Caspall, F. C. (1971). Error on choroplethic maps: definition, measurement, reduction. *Annals of the Association of American Geographers*, 61(2), 217–244.

MacEachren, A. M. A., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information ScienceInformation Science*, 32(3), 139–160.

MacEachren, A. M., Brewer, C. A., Pickle, L. W. (1998). Visualizing georeferenced data: representing reliability of health statistics. *Environment and Planning A*, 30, 1547–1561.

Reyes-Aldasoro, C. C., Bhalerao, A. (2006). The Bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition*, 39(5), 812–826.

Sun, M., Wong, D. W., Kronenfeld, B. J. (2014). A classification method for choropleth maps incorporating data reliability information. *The Professional Geographer*, 67(1), 72–83.

Sun, M., Wong, D., Kronenfeld, B. J. (2016). A heuristic multi-criteria classification approach incorporating data quality information for choropleth mapping. *Cartography and Geographic Information Science*, 1–13.

Xiao, N., Calder, C. A., & Armstrong, M. P. (2007). Assessing the effect of attribute uncertainty on the robustness of choropleth map classification. *International Journal of Geographical Information Science*, 21(2), 121–144.