AN EVALUATION OF CLASSIFICATION SCHEMES BASED ON THE STATISTICAL VERSUS THE SPATIAL STRUCTURE PROPERTIES OF GEOGRAPHIC DISTRIBUTIONS IN CHOROPLETH MAPPING

Robert G. Cromley and Richard D. Mrozinski

Department of Geography
University of Connecticut
Storrs, CT 06268-2148
USA

ABSTRACT

In choropleth mapping, most classification schemes that have been proposed are based on the properties of the data's statistical distribution without regard for the data's spatial distribution. However, one of the more important tasks associated with choropleth map reading is the task of regionalization and identifying spatial patterns. For this reason some authors have proposed class interval selection procedures that also consider spatial contiguity. This paper evaluates different classification schemes based on a data set's statistical as well as its spatial distribution. A comparison of the Jenks' optimal classification that minimizes within group variation and a contiguity based method that minimizes boundary error show that the latter method was not as strongly influenced by changes in the statistical distribution and produced a more complex map as measured by the number of external class boundaries present in the map display. (Keywords: data classification, choropleth mapping, spatial autocorrelation)

INTRODUCTION

Numerous classification methods for choropleth mapping have been proposed and evaluated (see Jenks and Coulson, 1963; Evans, 1977; Cromley, 1996). In general, most traditional and even optimal classification schemes such as the Jenks' optimal classification (Jenks, 1977) that minimizes total within group variation are based on the properties of the data's statistical distribution without regard for the data's spatial distribution. However, the task of regionalization is one of the more important tasks associated with choropleth map reading. Several authors (Monmonier, 1972; Cromley, 1996) have proposed class interval selection procedures that also consider spatial contiguity. The purpose of this paper is to evaluate classification schemes based on a data set's statistical distribution versus its spatial distribution. For this evaluation, the Jenks' optimal classification was chosen to represent schemes based on

statistical properties and Cromley's boundary error method (Cromley, 1996) was chosen to represent schemes that incorporate the spatial contiguity of the data values.

BACKGROUND

It has long been recognized that classification schemes have a major impact on the visualization of choropleth maps. Because the classification process transforms interval or ratio data into ordinal classes, information is lost converting individual algebraic numbers into ordinal classes. Secondly, grouping N unique data values into p different classes (N>p) implies that there are (N-1)!/(N-p)!(p-1)! different classification groupings. Monmonier (1991) has demonstrated how easy it is to distort the visual pattern of the data by manipulating the class interval breaks. The ambiguity caused by classification prompted Tobler (1973) to propose classless maps as an alternative to the classed choropleth map in which algebraic numbers are directly converted into graphic values. An areal table map (see Jenks, 1976) reduces this ambiguity even more by displaying the algebraic numbers directly within the outline of each area but visually recognizing patterns of spatial autocorrelation in geographic data sets would be more difficult.

To ensure that classification schemes try to represent the data distributions, different schemes have been evaluated within respect to how much error is associated with the classification (Jenks and Caspall, 1971) and the impact of class interval systems also have been analyzed with respect to the evaluation of pattern relationships (Monmonier, 1972; Olson, 1975; Dykes, 1994; Cromley and Cromley, 1996). While there are problems associated with any classification, well constructed classifications can aid the reader in most mapping tasks. Mak and Coulson (1991) found in perception tests that classed choropleth maps using the Jenks' optimal classification system (Jenks, 1977) were significantly better than classless maps for the task of value estimation although there was no significant difference in regionalization tasks.

The problem addressed here is to examine visually and quantitatively how well different classification schemes preserve the underlying spatial structure of the data. Cromley and Cromley (1996) found that quantile schemes frequently used in map atlases represented spatial patterns worse than classification based on minimizing the error associated with class boundaries. However, quantile classifications also generally produce worse representations than other classifications with respect to most statistical properties. The comparison here will be made between the Jenks' optimal classification and the boundary error method formulated by Cromley (1996).

Both of these methods are "optimal" in the sense that each minimizes or maximizes some performance measure. Both classification schemes are derived

from the same basic model. Based on Monmonier's work (1973) applying location-allocation models to the classification problem, Cromley (1996) has shown that optimal classification can be solved as a shortest path problem over an acyclic network. Using the number line associated with the sorted data values as an acyclic network, each arc connecting two nodes in the network represents a class interval containing data values. Given n points in the original data set, there would be n+1 nodes and n(n+1)/2 arcs in the acyclic network. For classifying data values in choropleth mapping, the cost value associated with each arc corresponds to an objective performance measure. By varying the definition of this performance measure, alternative optimal classifications can be constructed (Cromley, 1996).

Within the framework of this generic optimal classification model, the Jenks' optimal classification scheme defines the cost value for each class as the within class variation. By minimizing this value over all groups, the classification minimizes the total within group variation so that as much of the overall variation is "explained" by the classification as much as possible. The Jenks' optimal classification is also referred to as the VGROUP classification for the remainder of this paper.

Boundary error occurs whenever the boundaries between the classed areas on a map, referred to as external class boundaries, do not align with the major breaks in a three dimensional representation of the statistical surface (Jenks and Caspall, 1971). Classification should result in the boundaries lying within a group of contiguous area units, referred to as internal class boundaries, corresponding to minor breaks in the surface while the boundaries separating a grouping correspond to the major breaks in the surface. Within the generic model, the cost value for each class is now defined as the variation between the right- and left-hand area units associated with each internal class boundary. Only the deviations associated with boundaries separating area units within a class are counted while the deviations associated with boundaries separating area units belonging to different classes are not counted. By minimizing this cost value over all classes, the internal class boundaries should correspond to minor breaks in the surface and any regionalization should be fairly homogeneous. Because this classification (referred to as BGROUP) utilizes information regarding the relative location of data values, its implementation requires a topological data structure for the base map as well as the data values themselves.

DATA

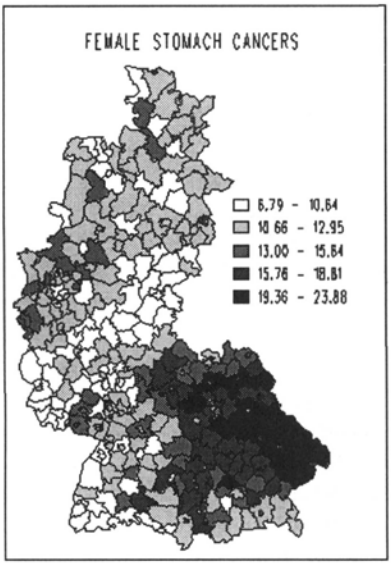
For evaluating these different approaches to classification, a cancer mortality data set was selected from West Germany originally published and mapped in *Atlas of Cancer Mortality in the Federal Republic of Germany* (Becker *et al.*, 1984). The data in this atlas were collected at the level of the **kreise**

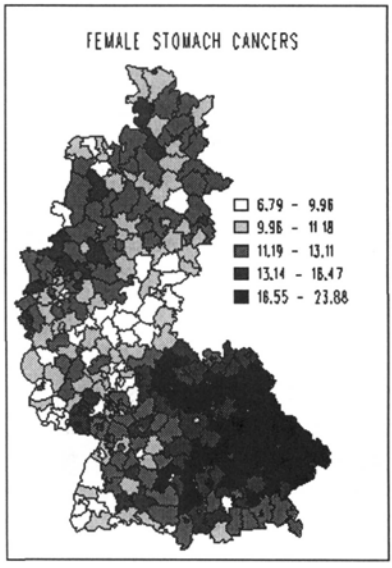
administrative unit for which mortality rates were estimated by the authors. Overall, there were 328 observations for each cancer; the **kreise** of West Berlin was removed from the original data because it was a detached unit and did not share common boundary with any other unit. Female stomach cancer, which was highly positively autocorrelated in West Germany, and ovarian cancer, which was randomly distributed over space, were chosen to test the effect of spatial arrangement on each classification. No negatively autocorrelated patterns were used because these patterns rarely occur in most geographic processes. Each of these cancer data sets also were slightly positively skewed in their respective statistical distributions.

In addition to mapping each cancer by both the Jenks' optimal classification scheme, VGROUP, and the spatial structure method, BGROUP (see Figures 1 and 2), three artificial data distributions were classified mapped for each cancer. These artificial data distributions are created to add differing levels of skewness in the statistical distribution for the same basic spatial arrangement of data values. A linear, arithmetic, and geometric progression (see Jenks and Coulson, 1963) of data values were generated and then assigned to **kreise** such that the ordinal position of each **kreise** was the same for each progression as for female stomach cancer and then the ordinal position of each **kreise** was the same for each progression as for ovarian cancer. Thus, each progression has the same statistical distribution for each cancer but a different spatial arrangement. Finally, to keep the number of maps to a manageable number, only a five class map was produced for each original cancer and every progression/spatial arrangement combination.

RESULTS

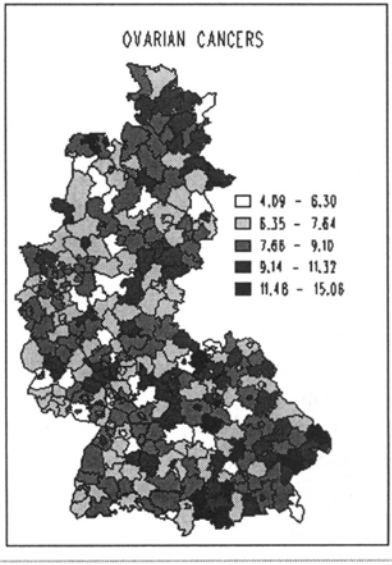
The Jenks' optimal classification of original female stomach cancer data was somewhat different than that for ovarian cancers as these two data sets had different statistical distributions although both were positively skewed (see Table 1). However, because the Jenk's optimal classification is based solely on the statistical distribution, the class intervals for the three progressions were the exactly the same for each progression regardless of how the values were spatially arranged. Secondly, the linear progression resulted in the same number of observations in each class. In this one case, optimal classification generates the same result as traditional quantile or equal interval schemes. Thirdly, as each artificial distribution became more positively skewed, more observations were grouped into the lower classes because the Jenks' classification is influenced by extreme values.

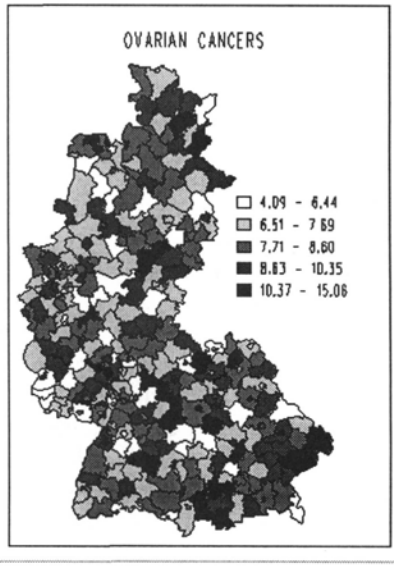




(a) VGROUP (b) BGROUP

Figure 1: Classified Female Stomach Cancers.





(a) VGROUP (b) BGROUP

Figure 2: Classified Ovarian Cancers.

The BGROUP classification method, in contrast, always produced a different set of class intervals for each statistical distribution/spatial arrangement combination (see Table 1). The number of observations for the linear progression that was positively autocorrelated (matched with the female stomach cancer arrangement) had fewer observations in the extreme classes than for the linear progression that was randomly arranged. As each progression became more positively skewed, more observations were grouped into the lowest data class although at a much lower rate than in the Jenks' optimal method.

TABLE 1
Number of Observations in each Class

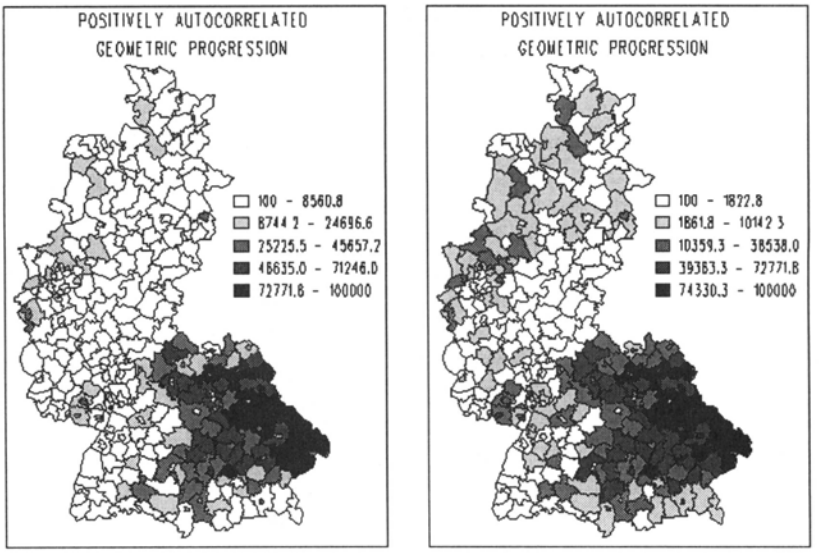
		Positively Autocorrelated		Random	
		VGROUP	BGROUP	VGROUP	BGROUP
Original Class Data*	#1	84	49	33	44
	#2	132	72	104	96
	#3	56	98	120	84
	#4	36	63	54	71
	#5	19	45	16	32
Linear Class Progression	#1	66	49	66	82
	#2	65	72	65	70
	#3	65	84	65	62
	#4	65	76	65	49
	#5	66	46	66	64
Arithmetic Class Progression	#1	119	103	119	84
	#2	67	59	67	71
	#3	53	57	53	66
	#4	46	63	46	63
	#5	42	45	42	43
Geometric Class Progression	#1	211	138	211	145
	#2	50	81	50	69
	#3	29	63	29	49
	#4	21	30	21	36
	#5	16	15	16	28

*The original data for the positively autocorrelated distribution were Female Cancers and the original data for the random distribution were Ovarian Cancers.

The overall result was that the spatial structure classification retained a higher level of visual complexity as the data distributions became more positively skewed especially for the data that was more positively autocorrelated (see Figures 3 and 4). In general, the visual complexity of a map increases as the spatial autocorrelation moves from high positive autocorrelation to random to high negative autocorrelation (Olson, 1975). The VGROUP classification of the geometric progression displayed a much larger homogeneous region of low values for the positively correlated distribution than for the spatially random distribution (see Figures 3a and 4a). Because the number of observations in each class was more balanced for the BGROUP classification than the VGROUP classification, higher level of visual complexity was retained for data set. For example, the large white area associated with the lowest class of Figure 3a is broken up into other classes in Figure 3b especially in the northern tier of **kreise**.

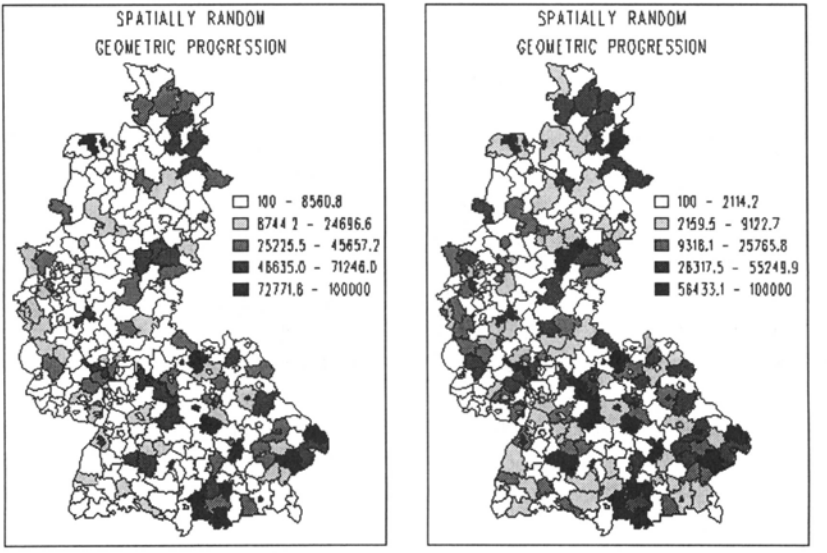
Quantitatively, this is measured first by the number of external and internal class boundaries generated by each classification. Because the boundaries between classes dominate the visual representation (Jenks and Caspall, 1971), the more external boundaries, the more visually complex the representation. In Table 2, the number of external and internal boundaries are matched against the Moran I coefficient for each data set. Regardless of the level of autocorrelation, the BGROUP classification always retained more external class boundaries than the VGROUP classification. Secondly, the BGROUP classification retained a similar number of external boundaries over the different data progressions.

Another measure of spatial autocorrelation and map complexity that has been used for map classifications is Kendall's tau (Monmonier, 1974; Olson, 1975). Similar to Moran's I index for metric data, Kendall's tau ranges from +1.0 to -1.0 for ordinal data with +1.0 associated with perfectly positive autocorrelation. With respect to Kendall's tau, the results were more mixed; for the positively autocorrelated distributions, the Kendall's tau value associated with BGROUP classification was higher for the stomach cancers data and the geometric progression and lower for the linear and arithmetic progressions. For the spatially random distributions, the Kendall's tau value was lower for the three progression and slightly higher for the ovarian cancers data. In general, the tau values were about the same for both classification with the exception of the positively autocorrelated geometric progression. Kendall's tau is influenced by an uneven number of observations in each grouping; the more uneven, the lower the value will be. Because the increasing skewness in the data resulted in the VGROUP's highly uneven number of observations in each class, the tau value is decreased.



(a) VGROUP (b) BGROUP

Figure 3. Classified Positively Autocorrelated Geometric Progression.



(a) VGROUP (b) BGROUP

Figure 4. Classified Spatially Random Geometric Progression.

TABLE 2
A Comparison of the Level of Spatial Autocorrelation
By Different Measures of Complexity

		Original Data*	Linear Progression	Arithmetic Progression	Geometric Progression
Positively Autocorrelated					
Moran I		0.759	0.636	0.723	0.807
# External Boundaries	VGROUP	466	515	483	322
	BGROUP	548	547	530	467
Kendall's Tau	VGROUP	0.461	0.485	0.493	0.381
	BGROUP	0.482	0.477	0.472	0.458
Random					
Moran I		0.080	0.053	0.069	0.095
# External Boundaries	VGROUP	600	664	599	426
	BGROUP	666	681	679	599
Kendall's Tau	VGROUP	0.036	0.036	0.050	0.036
	BGROUP	0.039	0.022	0.020	0.025

*The original data for the positively autocorrelated distribution were Female Cancers and the original data for the random distribution were Ovarian Cancers.

CONCLUSIONS

As expected, the Jenks' optimal classification was more strongly influenced by changes in the statistical properties of a data distribution than the classification that minimized boundary error. In all cases, the BGROUP classification resulted in a map display that had more external class boundaries than the Jenks' optimal classification. However, Kendall's tau measure for computing spatial autocorrelation for grouped ordinal data did not detect much difference between the two classifications except for the positively autocorrelated geometric progression. The overall result is that the BGROUP classification scheme probably retains more visual complexity and more homogeneous regions than the Jenks' optimal classification scheme.

REFERENCES

Becker, N., R. Frentzel-Beyme, and G. Wagner. (1984). *Atlas of Cancer Mortality in the Federal Republic of Germany*, Berlin: Springer-Verlag.

Cromley, R. (1996). A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographical Information Systems*, 10:404-424.

Cromley, E. and R. Cromley. (1996). An analysis of alternative classification schemes for medical atlas mapping. *European Journal of Cancer*, 32A:1551-1559.

Dykes, J. (1994). Visualizing spatial association in area-value data. Chapter 11 in *Innovations in GIS I*, M.F. Worboys (ed.), London: Taylor and Francis, 149-159.

Evans, I.S. (1977). The selection of class intervals. *Transactions of the Institute of British Geographers*, 2:98-124.

Jenks, G. (1976). Contemporary statistical maps -- Evidence of spatial and graphic ignorance. *The American Cartographer*, 3:11-19.

Jenks, G. (1977). *Optimal Data Classification for Choropleth Maps*. Occasional paper No. 2, department of geography, University of Kansas.

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

Jenks, G. and M. Coulson. (1963). Class Intervals for Statistical Maps. *International Yearbook of Cartography*, 3:119-134.

Mak, K. and M. Coulson. (1991). Map-user response to computer-generated choropleth maps: Comparative experiments in classification and symbolization. *Cartography and Geographic Information Systems*, 18:109-124.

Monmonier, M. (1972). Contiguity-biased class-interval selection: A method for simplifying patterns on statistical maps. *Geographical Review*, 62:203-228.

Monmonier, M. (1973). Analogs between class-interval selection and location-allocation models. *The Canadian Cartographer*, 10:123-131.

Monmonier, M. (1974). Measures of pattern complexity for choroplethic maps. *The American Cartographer*, 1:159-169.

Monmonier, M. (1991). *How to Lie with Maps*. Chicago: University of Chicago Press.

Olson, J. (1975). Autocorrelation and visual map complexity. *Annals of the Association of American Geographers*, 65:189-204.

Tobler, W. (1973). Choropleth maps without class intervals? *Geographical Analysis*, 5:262-265.