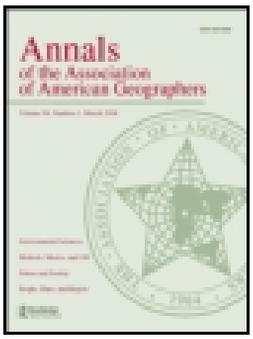


Routledge
Taylor & Francis Group

Annals of the Association of American Geographers

ISSN: 0004-5608 (Print) 1467-8306 (Online) Journal homepage: www.tandfonline.com/journals/raag20

Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series

Cynthia A. Brewer & Linda Pickle

To cite this article: Cynthia A. Brewer & Linda Pickle (2002) Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series, Annals of the Association of American Geographers, 92:4, 662-681, DOI: 10.1111/1467-8306.00310

To link to this article: https://doi.org/10.1111/1467-8306.00310

Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series

Cynthia A. Brewer* and Linda Pickle**

*Department of Geography, The Pennsylvania State University
**National Cancer Institute

Our research goal was to determine which choropleth classification methods are most suitable for epidemiological rate maps. We compared seven methods using responses by fifty-six subjects in a two-part experiment involving nine series of U.S. mortality maps. Subjects answered a wide range of general map-reading questions that involved individual maps and comparisons among maps in a series. The questions addressed varied scales of map-reading, from individual enumeration units, to regions, to whole-map distributions. Quantiles and minimum boundary error classification methods were best suited for these general choropleth map-reading tasks. Natural breaks (Jenks) and a hybrid version of equal-intervals classing formed a second grouping in the results, both producing responses less than 70 percent as accurate as for quantiles. Using matched legends across a series of maps (when possible) increased map-comparison accuracy by approximately 28 percent. The advantages of careful optimization procedures in choropleth classification seem to offer no benefit over the simpler quantile method for the general map-reading tasks tested in the reported experiment. *Key Words: choropleth, classification, epidemiology, maps.*

Choropleth mapping is well suited for presentation and exploration of mortality-rate data. Of the many options available, epidemiologists customarily use quantile-based classification in their mapping (see Table 5 in Walter and Birnie 1991). Our main goal was to evaluate choropleth classification methods to decide which are most suitable for epidemiological rate maps. The methods were evaluated by asking fifty-six subjects to respond to questions about individual maps and to make comparisons between maps in series. Seven classifications were compared in a two-part experiment involving nine series of U.S. mortality maps. Figure 1 presents an example of a test-map series. The four maps show death rates for white females from all causes, heart disease, all cancers, and stroke by health service areas for the conterminous United States.

The research follows from earlier collaborative work between Penn State and the National Center for Health Statistics (NCHS) previously published in the *Annals* (Brewer et al. 1997). For that project, we evaluated a series of choropleth color schemes in preparation for publishing the *Atlas of United States Mortality* (Pickle et al. 1996). The choropleth maps used in the earlier testing and in the atlas all used quantile-based classifications, following epidemiological practice. Use of quantiles in our 1997 color research and for the atlas prompted questions from cartographers and others about the wisdom of this classification decision. Thus, we set out to compare a range of classification methods in anticipation of increased production of mortality maps at the NCHS, the National Cancer Institute (NCI), and other health agencies through desktop geographic information systems (GIS) and Web resources. Classed one-variable choropleth maps are the most common in these and other mapping contexts (Mersey 1994). Therefore, we chose to test only these types of choropleth maps, though we will mention some alternatives, such as unclassed maps and bivariate maps, in our review.

Literature Review

The body of research on choropleth mapping has been reported in over seventy papers that span forty-five years of work. Unfortunately, much of the early in-depth research, conducted in the 1970s, is not easily uncovered, using modern search tools available on the Web or through university libraries, by researchers and analysts in other disciplines who are increasingly working with choropleth maps. Not surprisingly, these mapmakers are realizing the complexities of assigning classes to their data and are beginning to initiate their own investigations of choropleth mapping. Although psychologists and epidemiologists have recently conducted studies of a number of map-design elements, little or no research has been done on classification methods in these disciplines.

Annals of the Association of American Geographers, 92(4), 2002, pp. 662–681

Published by Blackwell Publishing, 350 Main Street, Malden, MA 02148, and 108 Cowley Road, Oxford, OX4 1JF, UK.

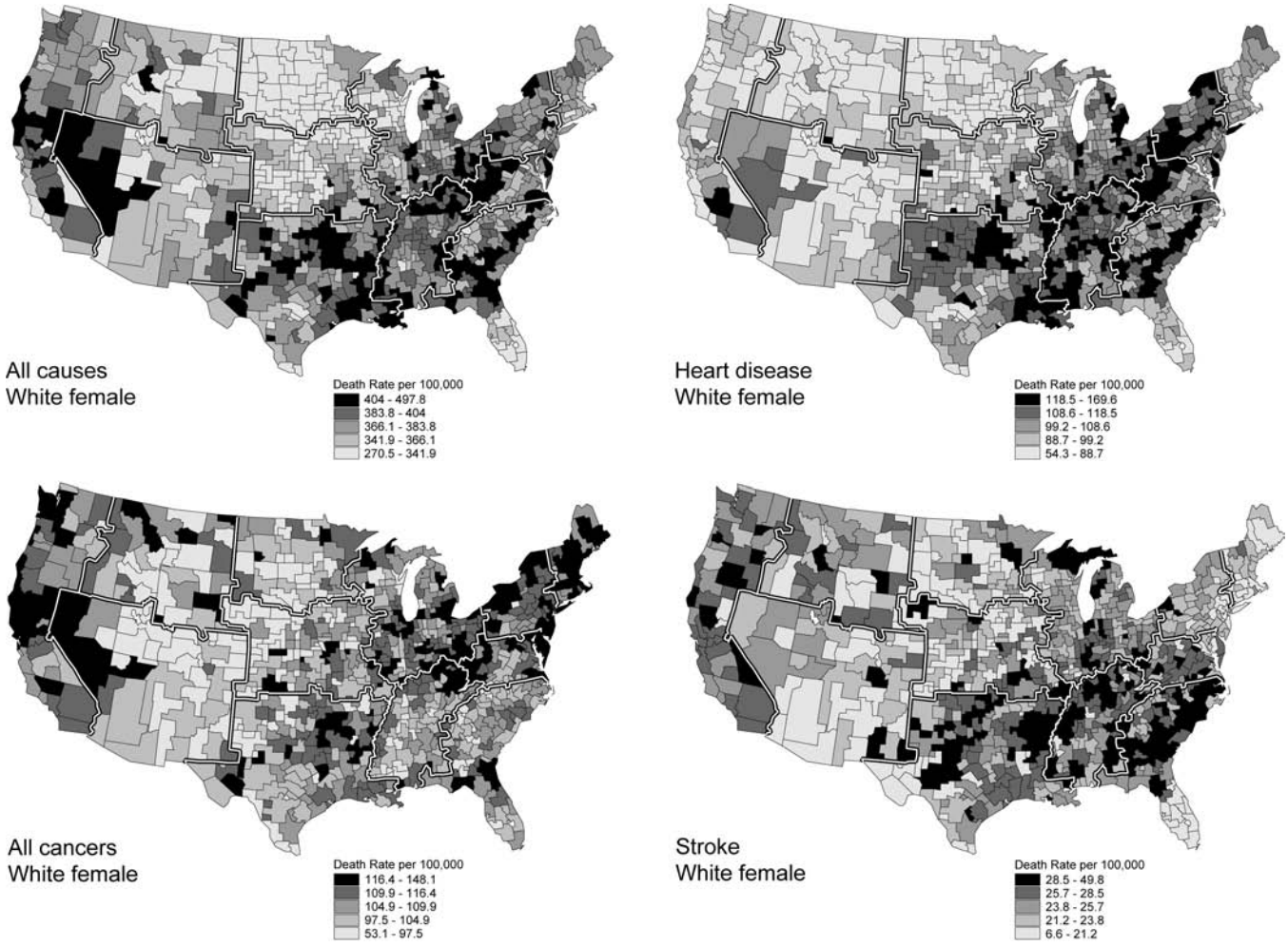


Figure 1. Example maps from part 1 of the experiment: quantile classification (QN) of series 3. Maps in the figure are shown in black and white and at 70 percent of size that subjects evaluated (test maps had yellow-green-blue color scheme).

Thus, a new review of the literature could be instrumental in bringing this early work back into circulation, to aid others in building on existing knowledge rather than repeating previous work.

Previously published reviews have focused primarily (though not exclusively) on methods suited to individual maps (Mackay 1955; Jenks 1963; Evans 1977; Paslawski 1984; Coulson 1987; MacEachren 1994; Robinson et al. 1995; Cromley and Cromley 1996; Dent 1999; Slocum 1999). General map-reading involves reading individual maps and comparing maps in series, and thus we include both of these aspects in our tests of map-reading accuracy. We begin our review with an emphasis on map comparison because it has not been systematically treated in previous reviews. Our focus on comparison is appropriate because epidemiological mapping offers crucial opportunities to compare spatial patterns in disease rates between races, between males and females, and through time, as well as encouraging comparison to patterns of potential causes. We complete our review with brief summaries of work on individual-map classification methods and other issues of choropleth map design.

Our review includes multiple references to "Jenks" classifications, so we include a short note here to aid readers in understanding this recurring theme. The group of classification methods proposed by George Jenks and his collaborators are generically referred to as "Jenks methods" or "optimized methods" (Jenks and Caspall 1971, Jenks 1977). The methods generally seek to minimize variation within classes. Coulson (1987) and Slocum (1999, 70–73) review Jenks methods, and Coulson notes that Jenks attributed the optimal classing algorithm to Fisher (1958). In contemporary GIS mapping tools, ESRI (Redlands, CA) uses Jenks calculations for their classification method labeled "natural breaks" (discussed further in Table 1 of the Methods section of this article).

Choropleth Map Comparison

Previous authors who reviewed classification methods included relatively brief recommendations on classification approaches specifically for map comparison. However, most of these recommendations were based on personal experience, not on objective experimental research. For example, Evans (1977) considered the goals of map comparison in just one page of his twenty-six-page paper. To examine absolute differences or change through time, he recommended using round-number class limits with equal intervals or perhaps geometric intervals that were shared by all maps. He also noted that comparison was facilitated by combining all data and calibrating the classification to the whole. The resulting variation in the number of classes on individual maps was deemed preferable to varying class limits among maps.

To examine differences in spatial pattern, Evans (1977) recommended that the same number of class intervals be derived from the data in the same way, such as quantiles, nested means, standard deviations, or natural breaks. With the exception of nested means, these classification methods are described in Table 1 below. Nested means (Scripter 1970) uses the mean as a middle class break and then divides ranges above and below the mean at their means. This process continues, producing classes symmetrically arrayed above and below a series of means. Evans felt that use of equal intervals was a poor choice for comparison because the approach was not related to central tendency. He also recommended treating zero as a separate class, especially when zeros were frequent, and then applying a classification algorithm to the remaining data range (or to each of the ranges above and below zero). Evans reported that map series usually use arbitrary classes that are not calculated using data characteristics so all maps can be compared.

Coulson (1987) recommended use of one of Jenks's methods in the construction of a map series. He noted, however, that a shared solution suitable for map comparison was suboptimal for individual maps. More recently, Cromley (1995, 1996) recommended exogenous classes for map comparison. Exogenous classes are determined using criteria relevant to the map topic but external to examination of the data distribution (e.g., use of 50 percent as a break for mapping two-party voting results; the winning party would be symbolized regardless of the statistical distribution of district results). Cromley's recommendation echoed an early warning from Dixon (1972) that data-based classifications (such as natural breaks, quantiles, and standard deviations) "inhibit" map comparison. MacEachren (1994) recommended use of natural breaks or Jenks optimization for choropleth mapping in general, but he noted that these approaches did not consider the need to compare maps. For comparison, he suggested applying the Jenks methods to the combined data range (consistent with Evans's [1977] recommendation). He also supported use of unclassed maps for comparison (discussed in the next section). He noted that quantiles and equal intervals highlighted differences in maps and that nested means were suitable for anchored comparisons if means were a relevant standard for the data.

In contrast to the review papers that only briefly mention map comparison, research has been conducted to investigate map comparison specifically. The most comprehensive research was conducted in the 1970s by Olson, Monmonier, Lloyd, Steinke, and Muller. Their research, reviewed below, focused on comparison of whole-map patterns evaluated using attributes such as correlation, correspondence, similarity, blackness, and complexity. The authors related these attributes to methods of classification and numbers of classes.

In early work, Olson (1972b) investigated the suitability of a sample of classifications for comparing maps. She constructed normal data distributions with twenty, forty, and one hundred units per map and classed these using eight methods based on quantiles, standard deviations, and nested means with two to five classes. She evaluated classifications by examining variability of rank correlations between pairs of classed variables for approximately 900 map pairs. She found that the choice of classification was more important when fewer units were mapped. Correlation estimates for the many map pairs were less variable with more classes and with approximately equal numbers of units in each class. For example, maps that were divided into more classes looked more similar regardless of the method used to produce the classes. She concluded that quantile classification was most effective in aiding map comparison.

Olson (1972a) extended the study to socioeconomic data that were not normally distributed (i.e., histograms for individual data distributions were not normal) and again compared variation in map-pair correlations. Standard deviation classing now had the lowest variability in correlations, and nested means also performed better than quantiles. In addition, increasing numbers of units did not decrease variability as they had before. This early work systematically and thoroughly tested one measure of map comparison accuracy. Olson found mixed results for quantiles and highlighted the importance of the relationship between the statistical distribution and the classification method.

In other early work, Monmonier (1972) agreed with Armstrong (1969) that data should be converted to standardized z-scores and classed by equal intervals to facilitate

visual comparison among mapped distributions. In addition, Monmonier (1975) proposed maximizing the visual correspondence between variables by using an optimal classification for the referent variable to which other mapped variables were compared. For example, a mortality-rate map could be the referent (dependent variable) to which a selection of behavioral variables would be compared. Intervals for related maps were then chosen to produce the greatest map similarity to the optimized pattern for the referent. Following a similar logic, Monmonier (1994) proposed minimum-change categories for dynamic choropleth maps, stressing the importance of stability of the image. He carefully selected breaks for two or three classes used during map comparison through a time series to assist identification of the largest changes.

Lloyd and Steinke (1976) found that classification method did not affect judgments of the similarity of choropleth maps when using five-class maps of normal data with equal-interval and Jenks classifications. Subjects were asked to select two of three maps that were most similar; subjects reported using blackness, complexity, and similarity of distributions to make their judgments. Based on these results, Lloyd and Steinke recommended holding "blackness" constant to aid map comparison. Their term "blackness" did not refer solely to areas that were solid black but, more generally, to use of equal land areas of each level of gray (percentages of black ink) used to symbolize classes on maps being compared. In a related paper, Lloyd and Steinke (1977) held blackness constant with an equal-area classification and concluded that similarity judgments with this classification were more comparable to calculated correlations between maps than for equal-interval or Jenks classifications. In 1981, Steinke and Lloyd extended their work with additional testing to conclude that cartographers should hold blackness constant for comparisons of both complexity and correlation. They (1983) also demonstrated that memorized images of choropleth maps and real maps were compared in the same way.

Lloyd and Steinke (1976, 1977) also acknowledged that an equal-area classification method could be programmed to hold areas of corresponding classes constant between maps, rather than requiring equal areas for all classes within maps. The work of Carr, Olsen, and White (1992) provided a later example of the equal-blackness approach. Carr and colleagues presented hexagon mosaic maps and determined classes using the area of map covered, with breaks at cumulative percents of 10, 25, 50, 75, 90, and 95. A less precise perspective is to consider quantiles as an approximation of an equal-area map when enumeration units are similarly sized.

Muller (1976) investigated the relationship between number of classes and choropleth map patterns. He examined four mathematically determined pattern attributes that may be used for map comparison: blackness, aggregation, complexity, and contrast. He used seventeen mapped variables with three to nine classes and calculated breaks using the Jenks-Caspall method (1971). The number of classes significantly affected calculated differences in the four pattern attributes, and attribute significance decreased as the number of classes increased. Variability of map patterns (differences in the four attributes) among different distributions also decreased with increasing numbers of classes. Muller (1976) concluded that use of few classes emphasized map pattern but that these three- and four-class maps often had markedly different patterns for the same dataset. More stable representations resulted with more classes, but these maps were less different from maps of other distributions, diminishing the generalization of pattern offered by the simpler maps.

In contrast to the approaches to map-pattern comparison taken by the researchers discussed above, Chang (1978) examined differences in map preference with quantile, equal-interval, arithmetic, standard deviation, and natural-breaks classifications. Subjects preferred simpler maps that had a small number of regions with a high degree of areal inequality among classes, low fragmentation and contrast, and high aggregation. These criteria produced higher preference ratings for arithmetic and equal-interval versions of simpler test maps and higher ratings for natural breaks with more complex test maps. These results were limited by lack of a map-reading accuracy component, but they provided a complementary view of criteria mapmakers may use in classification decisions.

In reports reflecting on atlas-making, Brewer (2001) and Becker (1994) described use of many classes that are shared among all maps in a series. Individual maps in an atlas may present data that range across a subset of the classes, but all maps will be directly comparable. For example, nearly half of the maps in a German cancer atlas would have five of the total of 20 classes developed for Becker's cancer-map series. In *Mapping Census 2000*, Brewer and Suchan (2001) used up to nine classes for map series, and some maps in these series used as few as four of these classes. Class breaks in *Mapping Census 2000* were developed using exogenous breaks, such as U.S. overall rates, combined with rounded arbitrary breaks (described in Brewer 2001). Becker's classifications were calculated using a square-root transformation of age-standardized rates.

Bivariate mapping offers an additional approach to comparison whereby two distributions are displayed simultaneously in a single choropleth map (e.g., Olson 1975b;

Carstensen 1986; Brewer 1994). The effectiveness of symbolizations for representing the overlay of two distributions has been a primary focus of bivariate mapping research. In contrast, Carr and Pickle (1993) suggested an approach that amalgamates rate data and numbers-of-people in a single classification strategy, similar to use of cumulative percentages of land area to set class breaks (Carr, Olsen, and White 1992). Carr and Pickle described determination of rate breaks using percent of people in each class. For example, the break for the highest class could be set when the accumulation of enumeration units with highest rates accounted for five percent of the population.

Other Choropleth Mapping Issues

Many general reviews of classification methods appear in the cartographic literature. Mackay (1955), Jenks (1963), and Jenks and Coulson (1963) presented reviews and critiques of early work. In addition to his discussion of map comparison, Evans (1977) reviewed sixteen systems for calculating class breaks, recommending attention to the overall shape of aspatial statistical distributions, such as frequency histograms. This strategy is still recommended in the classification overviews of current cartography textbooks and reference books (e.g., MacEachren 1994; Robinson et al. 1995; Dent 1999; Slocum 1999). Paslawski's 1984 review organized approximately twelve methods into a structured hierarchy. Coulson (1987) provided a review of classifications organized by user objectives and put particular emphasis on Jenks's methods. Cromley and Cromley (1996) also presented a classification review specific to medical atlas mapping.

A series of research articles on classification methods was preceded by the seminal contribution of Jenks and Caspall (1971) on optimal classification and map error, with tests of quantile, natural breaks, clinographic (related to percent area; Mackay 1955), and standardized methods. MacEachren (1985) tested relationships between map accuracy and characteristics of distributions and enumeration units. Smith (1986) stressed homogeneity within classes as a criterion for comparison of classifications to Jenks optimization. Many of the research papers reviewed in this and the preceding section tested Jenks classifications solely or as one of a few options in their map testing. Reflecting on this trend, MacEachren (1995, 384) suggested that many saw the maturation of optimal classification methods as the end of research on choropleth classification.

From a statistical perspective, but contrary to the optimal classification paradigm, Stegna and Csillag (1987) reported calculations showing that choropleth maps had maximum information content if all classes had equal frequency (quantiles). Their approach was to use iterative *t*-tests to decide the number of classes with statistically significant separability and then produce a quantile classification with this number of classes. Stegna and Csillag opined that users expected rounded class breaks and monotonically changing intervals; neither characteristic is a usual feature of optimal methods. Similarly, Paslawski (1983) proposed that users assumed constant intervals when reading choropleth maps.

Monmonier has published numerous papers proposing approaches that questioned, augmented, or offered alternatives to optimal classification. In 1972, he compared equal-interval and natural-breaks methods and suggested a hybrid method of shifting equal-interval breaks to the nearest natural break. In 1973, he investigated the similarity between classification and location-allocation problems. In work on pattern complexity (1974), he criticized the goal of internally homogeneous classes as shortsighted, given the greater importance of reducing complexity and enhancing pattern recognition in choropleth mapping. In 1982, he asserted that round-number class breaks were more easily remembered and promoted mental arithmetic, and therefore suggested use of rounding as an additional constraint on optimal solutions. He promoted the use of exogenous meaningful breaks and varied legend constructions, with the goal of useful accuracy. In addition to his 1975 and 1994 papers, discussed in the preceding section, this series of recommendations established Monmonier as a consistent critic of the optimal classification methods accepted by many cartographic researchers.

Recently, Cromley (1996) broadened the evaluation of optimized classification. He compared a variety of optimal classing criteria by testing four-class maps of constructed datasets of increasing skewness. He viewed the goals of map classification as removing spatial noise, highlighting spatial pattern, and enhancing positive or negative spatial autocorrelation. He tested classifications that minimized (1) distances to class medians, (2) within-class variation (Jenks), (3) distances to class midpoints, and (4) boundary error (Jenks [1963] introduced boundary-error indices for evaluating classifications). Cromley (1996) also compared these minimization methods to equal-interval and quantile classifications. He concluded that the best method would be a compromise produced by multiobjective programming that did best on all measures that were minimized by the four optimal methods tested.

Cromley's minimum-boundary-error method minimized within-class deviation between adjacent areas (so that boundaries mark major breaks in the statistical surface), maximized positive spatial autocorrelation, and was not affected by skewness in the distribution. Cromley and Mrozinski (1999) extended this method to classing ordinal data. The minimum-boundary-error method is one of

the few approaches that incorporate topology in the classification process. Other authors who have discussed the importance of spatial proximity in classification calculations include Monmonier (1972), MacDougall (1992), and MacEachren (1994).

Unclassed maps are another alternative to optimal choropleth classification. On these maps, lightness is proportional to the mapped rate: a map showing fifty-seven different data values would have fifty-seven different lightness values assigned to enumeration units. Tobler presented the original idea of unclassed maps in 1973 and was first rebutted by Dobson (1973). Investigation, application, and comment have continued in papers by Muller and Honsaker (1978), Muller (1979), Dobson (1980), Groop and Smith (1982), MacEachren (1982), Gale and Halperin (1984), Lavin and Archer (1984), Mak and Coulson (1991), and Kennedy (1994). Peterson's (1979) research included evaluation of classed and unclassed maps using a whole-map comparison task. He tested five-class maps produced with standard deviation classing and two versions of unclassed maps with different scalings for crossed-line shadings. He asked subjects to choose one of two maps that was most like, or most opposite to, a third map. He found little difference in subjects' judgments of correlations between maps and concluded that neither the generalization offered by classing nor the added information in unclassed maps was an advantage in the comparison of overall map patterns. In a recent investigation of unclassed choropleth maps, Cromley (1995) concluded that unclassed maps were too-many-class maps.

We would also caution that these too-many-class maps are, in effect, classed by the resolution of their output device. The perception of differences across the maps is controlled partly by the quality of the perceptual scaling in the color system used to assign printer or display colors to the myriad data values represented. For example, if the dark end of the lightness scale is compressed by a display device, differences between high data values will be less perceptible than difference between low data values. Simultaneous contrast (Brewer 1997a) also affects perception of lightness differences on unclassed choropleth maps (Muller 1979).

Issues of number of classes and complexity investigated by Muller (1976; see previous section) also came under discussion by other authors. Gilmartin and Shelton (1989) provided a review of the number-of-classes issue. MacEachren (1995) reviewed a method of selecting the number of classes based on evaluation of cognitive efficiency determined by the leveling out of Jenks's goodness-of-variance-fit measure. Cromley (1995) recommended setting a maximum class range to determine the minimum number of classes based on the overall data range. Others who have written about map pattern complexity without particular attention to the issue of map comparison include Olson (1975a), MacEachren (1982), Bregt and Wopereis (1990), and Mersey (1990).

The cartographic literature on choropleth mapping also includes work on color use, reliability, and dynamic mapping. McGranaghan (1989), Mersey (1990), Brewer (1994, 1996, 1997b), Brewer et al. (1997), and Olson and Brewer (1997) presented or tested recommendations for color use. MacEachren, Brewer, and Pickle (1998) examined methods of reliability representation for choropleth maps. Slocum, Robeson, and Egbert (1990), MacEachren and DiBiase (1991), Egbert and Slocum (1992), MacDougall (1992), Monmonier (1992), Slocum and Egbert (1993), and MacEachren (1995) all discussed dynamic choropleth mapping of varied types.

Cartographic research is shifting from communication (using maps to present known spatial patterns) to visualization questions (using maps to discover patterns). We expect that the role of classification beyond statistical optimization (MacEachren 1995, 384) and classification options in dynamic and interactive mapping of multiple variables will continue to be a focus of choropleth map research. Kraak and MacEachren (1999) summarize additional research opportunities in the emerging areas of visualization and dynamic mapping.

Experiment Methods

In order to test a sample of classification methods, we needed to prepare a sample of maps and a corresponding sample of questions about those maps. This section describes the choices we made as we constructed each of these samples, as well as describing the sample of map-readers we tested. The experimental designs used to balance classification methods and map series, so that all questions were asked for all methods, are described for both part 1 and part 2 of the experiment. Part 1 focused on comparing the accuracy of responses with different classifications. Part 2, a smaller experiment, examined the effect of using the same legend on all maps in a series.

Classifications

After evaluating the wide array of methods suggested in the literature, we selected seven methods of calculating class breaks for choropleth maps for our testing. Table 1 lists brief descriptions of each classification method tested. Methods are illustrated for a single distribution in Figure 2 with repeated small maps. The classifications we chose

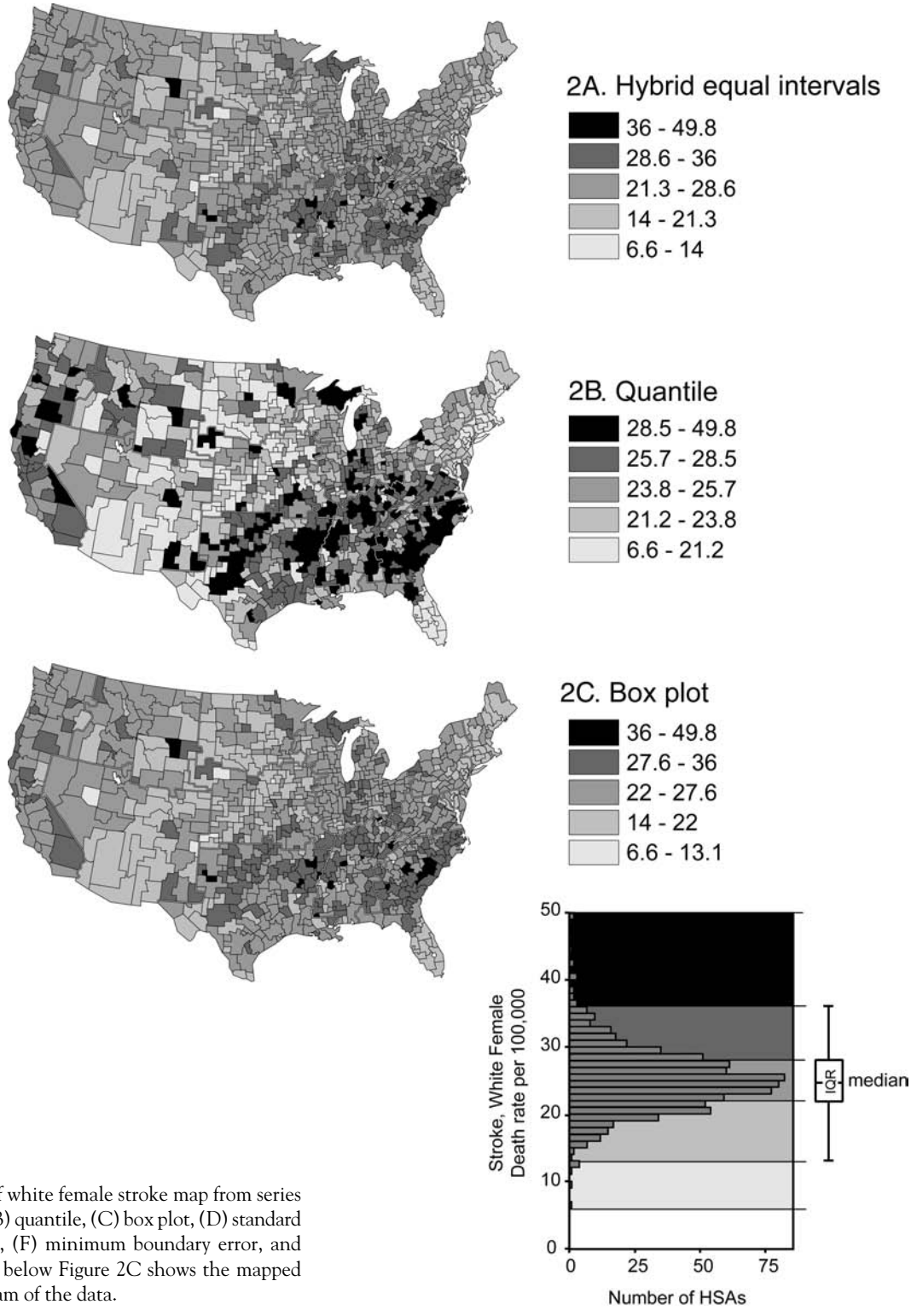
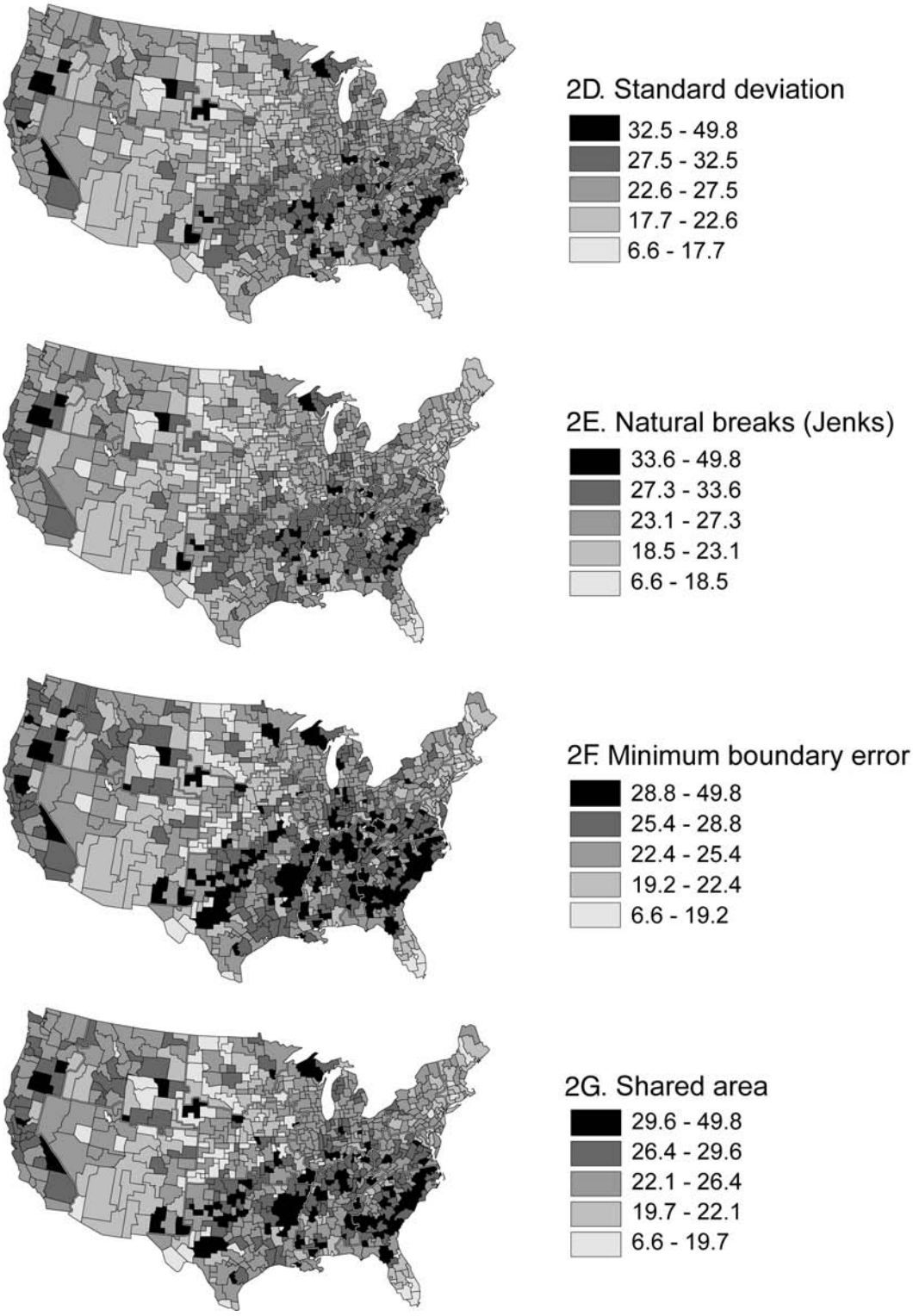


Figure 2. All classifications of white female stroke map from series 3: (A) hybrid equal intervals, (B) quantile, (C) box plot, (D) standard deviations, (E) natural breaks, (F) minimum boundary error, and (G) shared breaks. The graph below Figure 2C shows the mapped box-plot breaks with a histogram of the data.

were recommended by various authors for maps in series, as noted in the literature review above, with the exception of hybrid equal intervals and box-plot-based classes. These two methods were not previously proposed as appropriate for map comparison, but we felt they had potential worthy of testing. Additional methods, such as nested means, have been recommended for comparison, but we limited ourselves to seven methods to keep the project manageable in scope. We could have tested different numbers of classes or alternatives for selecting class breaks within



methods, such as putting a class break at the mean compared to having a class straddle the mean for standard deviation classing. Again, practical constraints on numbers of maps and subjects meant that we tested only a single version of each of the seven classification methods.

Use of a consistent calculation routine for all maps in the test left some maps with empty classes for some methods. For example, the many zero values in some data sets produced a second class with no values using the shared-area classification. Some maps with no outlier values or

Table 1. Summary of Classification Methods Tested

EI	The *hybrid equal interval* classification that we developed used the upper whisker of a box plot to define the highest category of outliers; see box-plot discussion below (BP) for explanation of whiskers. The remaining range of the data below the upper whisker was divided into equal intervals (e.g., equal steps of 7 deaths per 100,000). This approach was intended to be an improvement of the standard equal-interval method, which divides the overall data range into classes of equal range, regardless of the magnitude of extreme values. These extreme outliers are often present in epidemiological data, and they interfere with use of a regular equal-interval classification, making it an impractical or "straw man" method for mapping real data.
QN	The *quantile* method placed equal numbers of enumeration units into each class. With five classes, 20 percent of the units were in each class. Quantile classification is also known as percentile classification. With five classes, the test maps were quintile maps.
BP	The *box-plot*-based method had a middle class containing data in the interquartile range (the middle 50 percent of the data straddling the median). The adjacent classes extended from the hinges of the box plot to the whiskers, and the extreme classes contained outside and extreme values beyond the whiskers. Generally, the hinges of a box plot mark the top and bottom of the interquartile range, and the whiskers mark the last data values within 1.5 times the distance of the interquartile range above and below the hinges. For example, with an interquartile range of 10, from 33 to 43, the upper hinge would be as high as 58 (43 + 15). Data values higher than 58 and lower than 18 would be in the extreme classes for this example. See example map and corresponding box-plot and histogram in Figure 2C for a visual example. Box-plot-based classes were intended to be more suitable for skewed or asymmetric data distributions than a mean-based classification (see SD, below).
SD	The *standard deviation* classification had a middle class centered on the mean with a range of 1 standard deviation (0.5 standard deviation to either side of the mean). Classes above and below this mean class were also one standard deviation in range, from ±(0.5 to 1.5) standard deviations. The high and low classes contained remaining data that fell outside ±1.5 standard deviations.
NB	The *natural-breaks* method used was the implementation of the Jenks optimization procedure that was made available in ESRI's ArcView GIS software. In general, the optimization minimized within-class variance and maximized between-class variance in an iterative series of calculations. ESRI's documentation did not explain the specifics of their algorithm, but the ArcView natural-breaks method produced the same class breaks as did the Jenks algorithm that minimizes the sum of absolute deviations from class means (Terry Slocum, personal communication, e-mail, May 2000). See Slocum (1999) for a recent description of the Jenks algorithms.
BE	The *minimum-boundary-error* method used was also an iterative optimizing method (Cromley 1996). It was the only method tested that considered the spatial distribution or topology of the enumeration units (rather than their statistical distribution). In general, differences in data values between adjacent polygons were minimized in the same class and differences across boundaries were maximized between different classes (different colors). Larger differences in the data were, therefore, represented by color changes on the maps.
SA	The *shared-area* method used an ordered list of polygons ranked by data value to accumulate specific land areas in each class. With five classes, the extreme classes each covered ten percent of the map area. The middle class contained 40 percent of the area, and the remaining classes each contained 20 percent of the area. This method was based on the work of Carr, Olsen, and White (1992) and was intended to be a more sophisticated version of the constant-blackness (equal-area) method tested by Lloyd and Steinke in earlier work (1976, 1977). All maps in our series "share" the 10-20-40-20-10-percent area assignments, so we have labeled the method "shared area." We did not choose the previously used "constant area" or "equal area" terminology because classes within maps did not have equal areas.

with skewed distributions had four or even three classes when breaks were based on measures of central tendency (series 1, 2, 4, 5, and 6 for box-plot and standard-deviation classifications had empty lowest or highest classes; series are described in the next section). These empty classes were the result of classing real data, and we intended that they be included in our comparison of classification performance.

In addition, some of the methods were well suited to use of a diverging color scheme that would have accentuated differences from a mean or median class. We did not take advantage of these nuances, using the same color scheme for all methods. Generally, colors ranged from light yellow through green to dark blue and emphasized low-to-high sequences in the mapped data. Table 2 lists the approximate colors for the five- and seven-class maps. Each of the maps tested in part 1 of the experiment had five classes and each in part 2 had seven classes. More classes were used for these maps to compensate for the likelihood of empty classes on individual maps given the wider range in the matched legends.

Legend values on the test maps were rounded to tenths, and breaks were typically shared (e.g., 5.0 is

Table 2. Colors Used on Test Maps

	Munsell Notation	
Color Description	Hue	Value–Chroma
Dark purple (part 2 only)	7.5 PB	4–12
Dark blue	7.5 B	5–10
Medium blue-green	7.5 BG	6–8
Medium green	7.5 G	7–6
Light green	7.5 GY	8–4
Light yellow	7.5 Y	9–2
White (part 2 only)	N	10–

shared for classes 0.0–5.0 and 5.0–10.0), which allowed precise definition of breaks between very close data values. Class breaks indicated the values up to but not including the higher value in the class range. For example, classes of 0.0–5.0 and 5.0–10.0 would tell the reader that the first class included values between 0 and 4.9999 and the second class included values from 5.0000 to 9.9999. This is one of the standard methods of indicating class breaks when the mapped data have many more decimals than the mapmaker would like to show in the legend text. Some legends included information that revealed gaps between classes. For example, outliers in the top and bottom classes of box-plot legends were described using the range of actual values in the classes. For instance, 6.6–13.1 and 14.0–22.0 would indicate that the highest value in the low class was 13.1 and the lowest value in the next class was 14. These legend gaps provided additional information to the reader about the mapped data.

Table 3. Summary of Map Series Tested

Part One: Classification Test	
S1	WM, BM, WF, and BF lung-cancer mortality (pp. 48, 50, 52, 54)
S2	WM and BM HIV and unintentional-injury mortality (pp. 144, 146, 80, 82)
S3	WF all causes, heart-disease, all cancers, and stroke mortality (pp. 172, 36, 44, 76)
S4	WM motor vehicle, suicide, and homicide mortality (pp. 88, 120, 152) and percent urban
S5	WF breast-cancer mortality (p. 68), median income, percent of residents ages 25 and over who were college-educated, and percent urban
S6	WM heart-disease mortality at four time periods: 1982–1984, 1985–1987, 1988–1990, and 1991–1993
S7	WM stroke and lung-cancer mortality at two time periods each: 1979–1981 and 1991–1993
Part Two: Matched Legends Test	
S8	WM and WF liver-disease and chronic obstructive pulmonary diseases (COPD) mortality (pp. 136, 140, 196, 100)
S9	WM stroke mortality at four time periods: 1982–1984, 1985–1987, 1988–1990, and 1991–1993

Note: Page numbers identify corresponding maps in the *Atlas of United States Mortality* (Pickle et al. 1996), which present data from 1988–1992; time-series and socioeconomic maps did not come from the atlas. Abbreviations: W is for white, B for black, M for male, F for female; S1 for Map Series 1, S2 for Series 2, etc.

Map Series

Mortality datasets were selected for the experiment using maps in the *Atlas of United States Mortality* published by the NCHS (Pickle et al. 1996), with the exception of the time-series data, which were also provided by NCHS. Socioeconomic choropleth maps, chosen to complement selected mortality maps, were derived from 1990 census data.

Table 3 summarizes the series of map topics that were used in testing. The goal of part 1 was to evaluate the classification methods using accuracy of responses to questions about maps in series 1 to 7 (S1 to S7). The goal of part 2 was to evaluate the effect of using the same legend (i.e., exactly the same class breaks) on all maps in a series. Accuracy of responses to questions about maps in series 8 and 9 (S8, S9) were used in the part 2 analysis. Each series contained four interrelated maps. Figure 1 shows an example map series (S3 from part 1). The enumeration units used for the mortality maps in all series and for percent urban in S4 were health service areas (HSAs), which were aggregates of counties based on use of local hospital services (Pickle et al. 1996); there were 798 HSAs in the conterminous United States. The socioeconomic data in S5 were represented by county.

Subjects

People who participated in the experiment were each paid $10. All but one of the fifty-six subjects were undergraduate students at The Pennsylvania State University, and their majors were wide-ranging: engineering (twenty subjects), sciences (fifteen), liberal arts (eleven), education (four), and other majors (six). None of the subjects majored in geography. They ranged in age from 18 to 29 years, with a mean of 19.6. There were more males in the sample than females (forty-four versus twelve). The questionnaire took approximately forty-five minutes to complete. Students were recruited and tested in an on-campus residence building to produce a varied combination of majors within subject groups. We expect that the sample of students we tested adequately approximates the map-reading characteristics of interested map readers. A study of cluster identification using similar maps found little difference in performance between students and professional geographers and epidemiologists, except when subtle features of maps were important (Lewandowsky and Behrens 1995). Sampling students offered the advantage that subjects were practiced in using a multiple-choice testing format.

Test Questions

Test questions were multiple-choice and were designed to challenge subjects to read maps at all scales by asking about polygons (HSAs, seventeen questions total), regions (twenty-three questions), and whole maps (fourteen questions). Regions were approximate census divisions and were labeled with letters on an outline map

at the beginning of the test booklet. At each of the three question scales, subjects were asked about individual maps (twenty-eight questions) and were asked to compare maps (twenty-six questions). This set of questions follows Bertin's (1983) general classification of questions that can be asked of a graphic: data readout from a single area, patterns within a single graphic, and comparison of patterns across several graphics. This full range of questions is a necessary part of reading statistical maps, and thus we felt it was important to include simpler questions (such as comparing two HSAs) along with the more difficult questions (such as deciding which pair of maps in a time series represented the largest decrease in rates). Questions also differed in whether subjects needed to use the map legend to answer questions correctly. For example, subjects needed to compare legends and maps to decide which map in a series had the highest average rate. In contrast, a comparison of overall rate distributions on the maps did not require legend-reading, beyond understanding that dark-to-light represented high-to-low values. Table 4 lists example questions.

The level of precision required of subjects in their map-reading, and thus the extent to which this test was able to evaluate the nuances of the different classifications, is evident in the answer options offered with the example questions (see Table 4). We asked for broad evaluations of cluster locations and average rates. For example, we offered answer choices such as "higher than," "approximately equal to," and "lower than," rather than asking subjects to estimate specific averages. Understanding the types of questions asked in this evaluation of classifications is important for deciding whether the results provide appropriate guidance for selecting a method to produce maps for a specific use.

Correct answers for questions were calculated when possible, such as comparison of simple average rates in two regions. Answers were determined more subjectively for ten questions about clusters. Though this aspect of question construction added subjectivity to the results, it was important to also evaluate the maps with these types of questions because cluster interpretation is a key aspect of reading epidemiological maps. Another subjective aspect of these questions was individual subjects' interpretation of choices, such as "higher than" versus "approximately equal." Though we expected variation between subjects in this type of judgment, all subjects answered questions for all classifications, so bias in the answers of individuals was spread among the classifications. We also did not include questions that focused on detailed aspects of map-reading for which some classifications would be better suited, such as focusing on locations of extreme outliers.

Table 4. Example Test Questions

HSA question (legend needed)
(*S3*) For the marked area (Buffalo, NY), a possible rate for heart disease could be:
a) 151.7
b) 117.1
c) 105.3

Region question (legend not needed)
(*S7*) Fill in the blank. Rates for stroke in 1979–81 in Region B were _____ stroke rates in Region D.
a) generally higher than
b) approximately equal to
c) generally lower than

Map question (legend not needed)
(*S4*) Based on the overall pattern on the map, homicide rates are highest in what part of the country?
a) East
b) West
c) North
d) South

HSA comparison question (legend needed)
(*S1*) Fill in the blank. For the same area (Tallahassee, FL) the lung cancer rate for black males is _____ the lung cancer rate for black females.
a) higher than
b) similar to
c) lower than

Region comparison question (legend not needed)
(*S5*) Fill in the blank. Within Region G, areas with higher median incomes have _____ breast cancer rates.
a) higher
b) lower

Map comparison question (legend needed)
(*S9*) Which time span saw the largest decrease in the overall rate for the entire country?
a) 1982–84 to 1985–87
b) 1985–87 to 1988–90
c) 1988–90 to 1991–93

Our goal was to mimic questions that map-readers would ask of choropleth maps of real mortality-rate data. Though unambiguous questions would allow us to be more confident about the comparability of answers, they would be directed at only a small portion of the types of conclusions that people draw from maps. An appropriate range of questions about maps required inclusion of the ambiguity implicit in responses like "higher rates," "more clustered," "greater change," and so on. Our intention was that overall accuracy on many questions (forty-two questions in part 1) of varied types would indicate the overall performance with the classifications tested. We sought to evaluate classifications for maps in series in atlases suited to lay interests, so we tested maps with a wide range of questions and suitably generalized response options. In contrast, if a map author expects readers to

ask a particular type of question of a map series (such as studying deviations from means), then a classification suited to that narrow focus should be selected, regardless of its suitability for many other types of interpretations.

Inspiration for selection of map series and for comparison questions derived partially from the written summaries of mortality trends in the *Atlas of United States Mortality* (Pickle et al. 1996, 20–27). At the stage of selecting maps for series (Table 3) and in the very early stages of question construction, we saw these maps first as modified quantile maps, where quantiles were augmented by further dividing the lowest and highest classes. This modification converted five-class quantile maps to the seven-class maps seen in the atlas, with half as many HSAs in the first two and last two classes (percentages of HSAs in each class were 10, 10, 20, 20, 20, 10, and 10). In planning the experiment questions, we worked with all seven classifications of each map series, so we do not expect that this initial set of modified quantile maps from the atlas affected the questions. We acknowledge, however, that the initial quantile-based form of the maps could be a potential source of bias in our question-planning.

Table 5. Part 1 Classification and Series Combinations Seen by Each Subject Group

Page order	1	2	3	4	5	6	7
Group 1	SD S4	BP S3	NB S5	EI S1	QN S2	SA S7	BE S6
Group 2	EI S2	SD S5	SA S1	QN S3	BE S7	SD S6	BP S4
Group 3	SD S6	BE S1	SA S2	EI S3	NB S7	QN S4	BP S5
Group 4	QN S5	BP S6	EI S4	NB S1	BE S2	SA S3	SD S7
Group 5	QN S6	BP S7	EI S5	NB S2	SA S4	BE S3	SD S1
Group 6	NB S3	BE S4	QN S7	SA S5	BP S1	EI S6	SD S2
Group 7	BP S2	QN S1	SD S3	BE S5	NB S4	SA S6	EI S7

Group 1 example (explanation of first row of above table):

Classifications (see Table 1)	**Series** (see Table 3)
Standard deviation, SD	S4, vehicle, suicide, homicide, urban
Box-plot, BP	S3, major causes
Natural breaks, NB	S5, breast cancer, education, income, urban
Hybrid equal interval, EI	S1, lung cancer, B/W M/F
Quantile, QN	S2, HIV, injury, B/W
Shared area, SA	S7, stroke and lung cancer time-series
Minimum boundary error, BE	S6, heart disease time-series

Note: Each subject saw each classification and each series once in seven combinations (they did not see all possible combinations of classifications and series).

Experimental Design

The experiment was divided into two parts. The majority of the testing, part 1, was designed to compare map-reading accuracy for seven classification methods. Part 2 was a smaller portion of the experiment that evaluated map-reading accuracy with matched legends for maps in series.

Part 1: Classification Testing. For part 1, each subject saw each map series and each classification once. Series and classification were ordered randomly for seven groups of eight subjects (Table 5). There were six questions per map, for a total of forty-two questions. Altogether, fifty-six subjects completed the test, for a total of 2,352 observations. This design had a power to detect a true 10-percent difference in overall accuracy between classification methods 75 to 80 percent of the time.

Part 2: Matched Legends Testing. Each subject saw two map series in part 2 (S8 and S9 in Table 3). The versions of the series that an individual subject saw had the same classification (hybrid equal intervals or natural breaks—EI and NB in Tables 6 and 1). One of the two series examined by subjects was presented with matched legends based on hybrid equal intervals or natural breaks (EI-M and NB-M in Table 6). Classifications for matched legends were calculated by merging all four datasets in a series and determining breaks using this aggregated dataset. Figure 3 presents the example of S8 maps with matched legends (EI-M).

Hybrid equal intervals (EI and EI-M) and natural breaks (NB and NB-M) were tested in this part to give some variety of classification methods, but they were not the primary aspect of investigation in this smaller test. They were better suited to lumping the data into one dataset for classification calculations than some of the other methods, such as minimum boundary error. Again, there were six questions per map, yielding twelve questions total. Forty-eight of the fifty-six subjects participated in part 2. A subset of subjects was used to maintain a balanced number of twelve subjects in each of four groups, for a total of 576 observations. Subjects answered questions for parts 1 and 2 during the same test session, with part 2 questions on pages 8 and 9 of the test booklet. This design had a power of over 95 percent to detect a true 15-percent difference in

Table 6. Part 2 Classification and Series Combinations Seen by Each Subject Group

Group 1′	EI-M S8	EI S9
Group 2′	EI S8	EI-M S9
Group 3′	NB-M S9	NB S8
Group 4′	NB S9	NB-M S8

Note: Codes for classification methods are given in Table 1, with the addition of EI-M, the hybrid equal interval method with matched legends, and NB-M, the natural breaks method with matched legends. An explanation of how to read this table is given with Table 5.

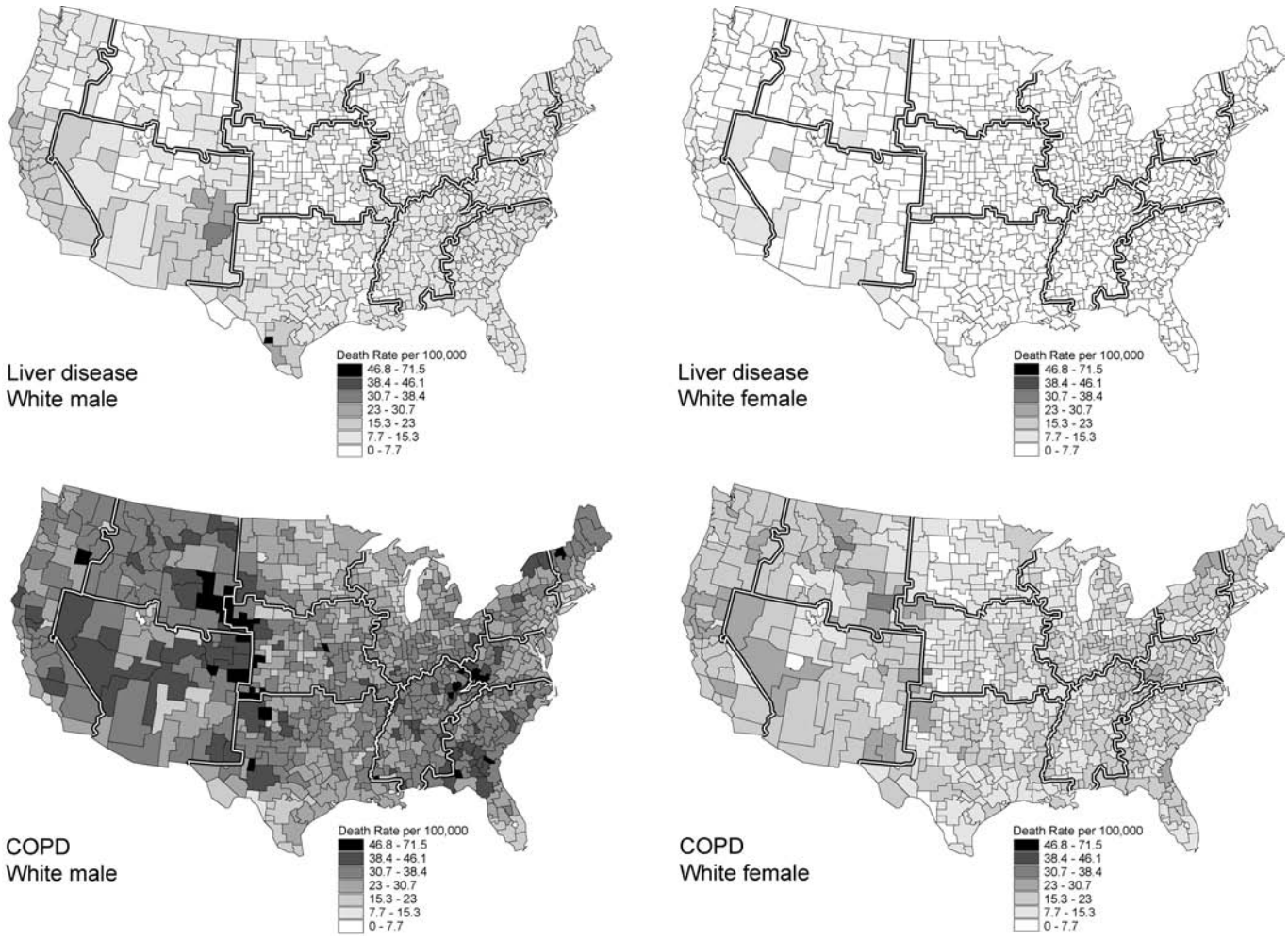


Figure 3. Example maps from part 2 of the experiment: hybrid equal interval classification (EI-M) of series 8 with matched legends. Maps in the figure are shown in black and white and at 70 percent of size that subjects evaluated (test maps had yellow-green-blue-purple color scheme).

proportion accurate between the matched and unmatched legend series.

Statistical Methods

Logistic regression analysis was used to test the differences in accuracy of response using the seven choropleth classification methods (part 1) and using matched versus unmatched legends (part 2). Results from parts 1 and 2 were analyzed separately. Logistic regression is used to model the relationships between a set of predictor variables and a dichotomous dependent variable (in this case, correct and incorrect responses) (Cox 1970). We made our logistic calculations using SAS PROC LOGISTIC. For the part 1 calculations, the quantile method was considered the referent method against which the other methods were tested, because it ranked first in overall observed accuracy.

In addition to classification method and legend matching (for part 2 only), we also included task type, legend use, symmetry, and map series (Table 3) as predictor variables in the model. These additional variables (described below) were included because we wanted to account for classifications that might produce particularly good or poor performance on a subset of questions or maps.

Task type had six levels reflecting the scale of the question and whether the question required comparison across maps (as described in the Test Questions section above). Questions were classed into tasks that required within-map decisions for (1) HSAs, (2) regions, and (3) whole maps and into tasks that required map comparisons at these scales, referred to as (4) HSA-comparison, (5) region-comparison, and (6) map-comparison tasks. Subjects examined a single map for within-map questions and two, three, or four maps for comparison questions. For example, subjects were asked to compare average

rates in two regions within one map for a region task and to compare average rates in the same region on two different maps for a region-comparison task.

Additional variables included legend use and symmetry. Legend use was a dichotomous variable that coded whether subjects needed to use the legend to answer a question correctly. Another dichotomous variable was included that coded whether the data distribution for the map was symmetric or asymmetric. Judgments of symmetry were based on the appearance of a frequency histogram for each mapped variable. This variable was included because some classifications might be less effective when the mapped data were skewed or had extreme outliers.

Interactions between all pairs of these variables were included in the full LOGIT model, and a backwards elimination process was used to arrive at a minimum set of main effects and interactions that produced predictions of accuracy that were not significantly different than the full model. Main effects remained in the model if they were significant at the 0.05 level, while interaction terms were required to meet a stricter criterion of 0.01-level significance to be included in order to reduce the number of spurious significant results among the many tests of possible interactions. Goodness-of-fit of the model was tested by likelihood ratio tests.

Results

Part 1: Response Accuracy for Classifications

Subjects were most accurate using the quantile method of classification (75.6 percent overall), followed by the minimum-boundary-error method (72.6 percent). Natural breaks (Jenks) and hybrid equal intervals produced similar accuracies (69.9 percent and 69.4 percent). The methods that yielded the poorest accuracies were standard deviation (67.6 percent), shared-area (66.4 percent), and box-plot (64.6 percent). Table 7 lists observed accuracy for each classification for the six task types.

The final LOGIT model for part 1 included classification, map series, task, legend use, task by legend use, and task by series (Table 8 lists degrees of freedom, chi-square values, and *p* values for each of these effects). This model was produced through backwards elimination, and model estimates of accuracy were not significantly different from those of the full model. From this simpler model, we concluded that there were significant differences in the accuracy of question response associated with classification. Task also affected accuracy, because some question types were more difficult than others (see Table 7). Task interacted with series because some tasks were more difficult than others for some series. Task interacted with legend use primarily because questions requiring region- or map-comparison with simultaneous comparison of legends were particularly difficult. Series and legend use were also significant main effects because questions for some series were more difficult and questions for which legend use was required were more difficult.

These additional main effects and interactions involving task, series, and legend use were important because they accounted for variation in accuracy that was not explained by differences in classification method alone. A crucial aspect of this model was that classification did not significantly interact with any other predictor variable. Thus, classification method produced consistent differences in accuracy regardless of characteristics of the question (differences in tasks and legend use) and of the maps (series and symmetric data).

Table 9 shows the predicted percent accuracy for each classification method by type of task, averaged over the

Table 7. Part 1 *Observed* Percent Accuracy by Classification Method for Question Types

	Task					
	Within-Map Questions			Between-Map Questions		
Classification Method	HSA $n = 72$	Region $n = 72$	Map $n = 24$	HSA-comparison $n = 32$	Region-comparison $n = 72$	Map-comparison $n = 64$
Quantile	90.3	84.7	70.8	84.4	62.5	60.9
Minimum boundary error	87.5	81.9	83.3	81.3	47.2	65.6
Natural breaks (Jenks)	84.7	70.8	62.5	78.1	58.3	64.1
Hybrid equal interval	83.3	79.2	70.8	71.9	54.2	59.4
Standard deviation	79.2	80.6	75.0	75.0	43.1	60.9
Shared area	86.1	75.0	62.5	43.8	56.9	57.8
Box plot	81.9	69.4	79.2	84.4	51.4	39.1

Note: n is the number of responses for which percentages are calculated; e.g., $n = 72$ is eight subjects responding to nine questions for each classification. In part 1, twenty-one within-map and twenty-one between-map questions were asked for each classification.

Table 8. Part 1 Summary of Effects in Final LOGIT Model

Effect	*df*	Wald Chi-square	*p*
Classification	6	18.630	0.0048
Task	4	15.741	0.0034
Legend use	1	10.684	0.0011
Series	6	95.392	<0.0001
Task by legend use	3	20.874	0.0001
Task by Series	21	240.198	<0.0001

Notes: *df* = degrees of freedom. *p* = probability; *p* values are less than .01 for each effect in the final model. These significance levels indicate that each remaining effect explains enough variation in the data that their removal would produce a final model with significantly poorer predictions than the full model.

series of test maps and variation in legend use and symmetry. The quantile method ranked best for all tasks. Despite small numbers for individual classification and task combinations, this pattern is generally true for the observed accuracies as well (Table 7). The quantile method had a consistently high predicted accuracy for all tasks, even when responses with other methods fell below 60 percent accuracy on more difficult tasks (region and map comparison).

We also computed the odds ratio of a correct response for each of the classification methods relative to quantiles, accounting for other effects in the model involving task, legend use, and series. Quantiles was chosen as the referent for these calculations because it produced the most accurate responses, and thus it had an estimated odds set at 1.0. Figure 4 provides a graphic representation of odds ratios and confidence limits for all classifications relative to quantiles. Though minimum-boundary-error classification produced responses 80.8 percent as accurate as quantiles, its upper 95-percent confidence limit exceeded 1.0 substantially (95 percent confidence limit = [0.539, 1.212]). This overlap is reflected by the lack of a significant difference between minimum boundary error and quantiles reported with probabilities for pairwise comparisons in Table 10 (p=0.303). Estimated accuracy for natural breaks was 67.4 percent of quantiles, and the upper confidence limit for natural breaks barely overlapped the referent by 0.006 (95-percent confidence limit = [0.451, 1.006]). This small difference is echoed by the relatively large p value of 0.053 for the quantiles and natural breaks pair in Table 10 (p barely misses significance at our threshold of 0.05). Hybrid equal intervals had an estimated accuracy of 66.0 percent, which was similar to that of natural breaks. The upper 95-percent confidence limit for hybrid equal intervals was just below 1.0 (95-percent confidence limit = [0.442, 0.986]). The remaining methods (standard deviation, shared area, and box plot) have unambiguously poorer estimated accuracies of 57.6 percent, 53.3, and 47.6 percent relative to quantiles. The upper 95-percent confidence limits for these methods each fell short of 1.0 by at least 0.14.

Part 2 Results: Response Accuracy for Matched Legends

Overall, the use of matched legends led to an improvement in response accuracy (79.9 percent observed accuracy for matched legends, compared to 70.5 percent for unmatched legends). The logistic analysis, however, showed striking differences by task, even after controlling for other significant factors (classification method, map series, use of legend, symmetric data distribution). Region- and map-comparison tasks were harder than within-map HSA and region questions (the twelve questions for part 2 included only these four task types). Matched legends significantly improved performance for comparison questions, but they did not improve performance for within-map questions. The estimated odds ratio showed a 4-to-1 improvement in accuracy (95-percent confidence limits = [2.22, 7.22]) when matched legends were used for the comparison questions but a slight non-

Table 9. Part 1 *Predicted* Percent Accuracy by Classification Method for Question Types

	Task					
	Within-Map Questions			Between-Map Questions		
Classification Method	HSA	Region	Map	HSA-comparison	Region-comparison	Map-comparison
Quantile	87.1	82.4	78.9	81.1	61.3	67.1
Minimum boundary error	86.0	80.0	75.6	77.8	57.3	62.8
Natural breaks (Jenks)	85.0	77.8	72.6	74.8	53.8	58.9
Hybrid equal interval	84.8	77.6	72.2	74.4	53.5	58.5
Standard deviation	84.0	75.8	69.9	71.9	50.9	55.5
Shared area	83.5	74.8	68.5	70.5	49.5	53.8
Box plot	82.7	73.2	66.5	68.2	47.4	51.3

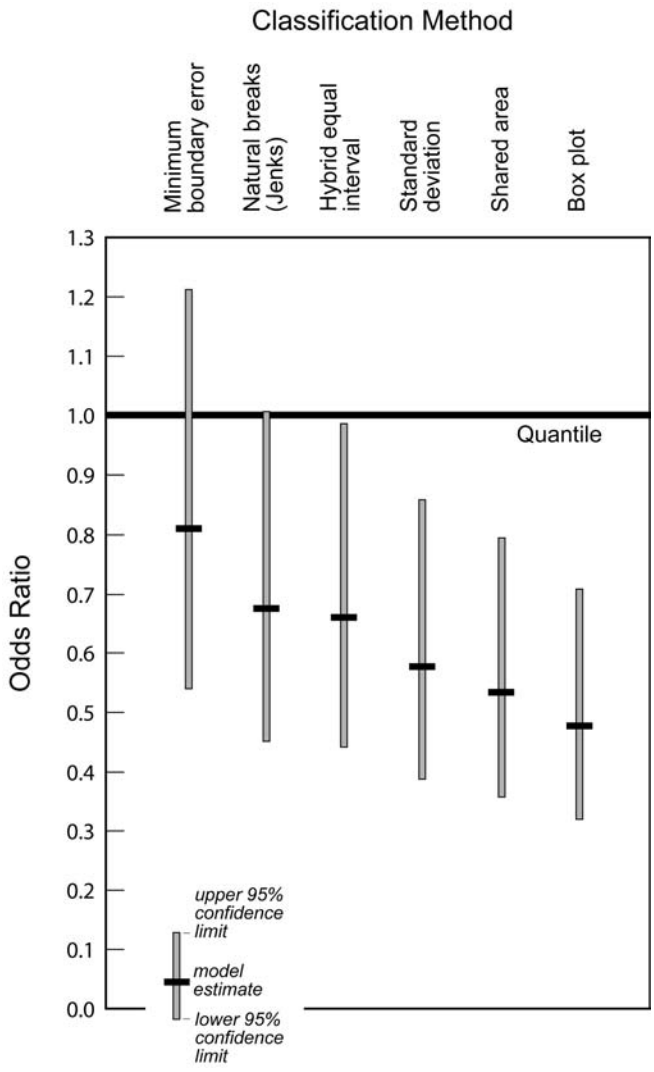


Figure 4. Odds ratios for percent accuracy for classification methods relative to accuracy with the quantile method. As shown in the lower-left key, black horizontal bars represent model estimates of overall accuracy for each method. Upper and lower 95-percent confidence limits are represented by the extent of the vertical gray bars. Methods are ordered from most to least accurate; quantile ranked first in accuracy. Bars for minimum boundary error and natural breaks (Jenks) overlap the 1.0 estimate for quantile, the referent for this calculation, and were thus not significantly different.

significant decline in accuracy for the within-map questions (odds ratio = 0.64, 95-percent confidence limits = [0.31, 1.30]). These modeled results were consistent with the observed accuracy percentages: 71.7 percent of the comparison questions were answered correctly when the legends were matched, compared to only 43.3 percent accuracy without matched legends (an increase of 28.4 percent). For within-map HSA and region questions, these percentages were 85.7 percent and 89.9 percent, respectively.

The final models for part 2 were run separately for within-map and comparison tasks to further examine the benefits of using matched legends when comparing maps. Because of the limited number of maps and tasks used for part 2, we had insufficient data to examine both types of tasks in a single analysis. Additional significant main effects in these reduced models were map series for the comparison tasks and legend use, map series, and symmetric data distribution for the within-map tasks. Subjects were more accurate with the hybrid equal-interval classification than natural breaks (regardless of whether or not legends were matched), which was a reversal of the order seen in part 1 for these two classification methods. This difference was significant for the comparison tasks but not for the within-map tasks. The more limited set of twelve questions in part 2 (compared to forty-two in part 1) produced a less compelling comparison of classification methods.

Additional significant results from the part 2 analysis reveal characteristics of within-map questions and map series, but they do not affect conclusions about the benefits of matched legends. Within-map questions were significantly harder when legend use was required for an accurate response ($df = 1$, chi-square = 4.1, p value = 0.04) or when the data were symmetrically distributed ($df = 1$, chi-square = 21.0, p value < 0.0001). The latter result was more an outcome of our design of more challenging questions for these maps than a generalized difficulty with symmetric distributions. Accuracy using the liver-disease maps (S8) was significantly worse for the within-map tasks but significantly better for the comparison tasks compared to the results using the stroke maps (S9). Again, this result was more an outcome of the difficulty of the particular questions asked than a generalized difficulty of S8 and S9 for these different tasks. As we designed question sets for each series, we did not seek equal difficulty between map series because all subjects saw all series and all questions.

Conclusions

Classification methods best suited for choropleth maps in series intended for a wide range of map-reading tasks were quantiles and minimum boundary error. Natural breaks (Jenks) and hybrid equal intervals were not meaningfully different from each other in accuracy of responses and formed a second grouping in the results; they produced responses approximately 67 percent as accurate as quantiles (Figure 4). A third grouping in the results was standard deviation, shared-area, and box-plot classifications, which all produced responses less than 60 percent as accurate as quantiles.

Table 10. Pairwise Comparisons of Classification Methods

	Quantile	Minimum Boundary Error	Natural Breaks (Jenks)	Hybrid Equal Interval	Standard Deviation	Shared Area
Minimum boundary error	0.303					
Natural breaks (Jenks)	0.053	0.365				
Hybrid equal interval	0.042*	0.315	0.921			
Standard deviation	0.007**	0.090	0.429	0.489		
Shared area	0.002**	0.037*	0.237	0.279	0.695	
Box plot	0.0002**	0.008**	0.078	0.096	0.330	0.560

Note: Significant differences in accuracy of response for pairs of classification methods are represented by * p values < 0.05 and ** p values < 0.01. Methods are listed in order of overall accuracy.

These results may surprise cartographers. Quantile classification is one of the simplest methods, and it produced accuracies not significantly different from or better than two of the most sophisticated optimal methods (Jenks natural breaks and minimum boundary error) when subjects answered a wide range of questions about series of epidemiological maps. Cartographers have long criticized the widely varied class intervals produced by quantiles. The method often produces an extreme range of values in the highest class for socioeconomic data (which is often skewed to include more low values).

Concern by cartographers about the choice of a modified quantile method for the *Atlas of United States Mortality* (Pickle et al. 1996) was an initial motivation for this research. Epidemiologists have long used quantiles in mapping with little question that they are an appropriate approach (see Walter and Birnie 1991, table 5). Epidemiologists value characteristics offered by quantile classifications; classes are usually centered on the median (a robust indicator of central tendency) and they systematically group enumeration units above and below the median into classes with equal frequencies regardless of their relative values. Perhaps more importantly, mapped epidemiologic data typically consist of age-adjusted disease rates, and these values are only meaningful in relation to other similarly adjusted rates. That is, the rank of an enumeration unit is more meaningful than the actual value of its rate. Epidemiologists are most familiar with these sorts of data, and their instincts about appropriate classifications of their data for mapping seem to be correct.

We consider the results of this study to have applicability beyond epidemiological mapping. There are hints throughout the cartographic literature on choropleth map comparison that suggest that quantiles perform well. In her early controlled tests, Olson (1972b) found that they produced comparisons most similar to correlations. They approximate equal-area classifications (with a simpler algorithm) when units are similar in size, and Lloyd and Steinke (1976, 1977) showed that this equal-blackness property fostered accurate map comparisons. Both Olson (1972b) and Muller (1976) also found that classification mattered less with increasing numbers of classes and increasing numbers of enumeration units. Stegna and Csillig (1987) suggested that quantile classes contained maximum information, and Evans (1977) recommended their use for comparing maps in an early review. More recently, Slocum (1999, 195–96) has also supported their use for map comparison. Nevertheless, cartographic researchers have also shown quantiles to be less effective for some map-reading tasks, and thus these results were unexpected.

Comparison tasks make up only part of the challenge for quantiles in this experiment. Subjects were also asked to respond to tasks that required estimating averages for regions or comparing averages between regions within a map. These are tasks for which the unpredictable ranges in quantile classes should have hindered response accuracy, but they did not. Perhaps the relatively large number of polygons in the highest class, compared to other methods, provided a sort of psychological scaling that compensated for underestimations of aggregate rates and therefore improved accuracies (suggested by Alan MacEachren, personal communication, conversation, August 1999). Perhaps quantiles provided greater visual contrast between regions that assisted in map-reading. We suggest this because the minimum-boundary-error method should also create better regionalization, and perhaps this characteristic caused both methods to produce better accuracies.

Quantiles can also be thought of as converting the mapped data to ordinal rankings, and perhaps this level of understanding is well suited to the general map-reading we asked of test subjects. As noted above, this quality is particularly relevant to epidemiological mapping of age-adjusted rates or SMRs (standardized mortality ratios—another version of adjustment), both of which are artificial constructs. The primary meaning of these numbers is

in relation to other like numbers, and from this perspective ranking provides a suitable visual summary. If rates are adjusted in two different ways to produce two maps, the ranking of places and thus the quantile classing will remain fairly constant, but equal-interval breaks, for example, may be quite different using the two methods of adjustment.

The group of methods with the poorest accuracy each had the characteristic that the middle class contained a large proportion of the data values. The worst one, box-plot classes, had a middle class with 50 percent of the data values. The shared-area method placed 40 percent of the mapped area into the middle class. Standard deviations would contain 38 percent of the data values in one standard deviation centered on the mean (from −0.5 to +0.5 standard deviations) for normal distributions and contained an even larger percentage for skewed distributions. Each of these methods produced finer differentiation of the extreme values at the expense of a lack of overall differentiation on the maps. Since this is a characteristic common to each of these maps, we suspect that this is the primary reason for the poor accuracies with the widely varied questions we tested. Questions about extreme values would perhaps have been answered more accurately with these methods, but they comprise only one of many types of questions we should expect to ask of maps in series. Any class with a disproportionate number of map units, regardless of the similarity in their data values, may hamper map-reading. It also follows that these methods could be adjusted to provide more breaks within the middle class (e.g., adding a break at the median for the box plot) to produce better map-reading using the same classification logic.

The results also show that subjects found it more difficult to perform some of the map-reading tasks asked of them than others. Interpreting broader mapped rate distributions and comparing maps were difficult tasks, as were questions requiring legend-reading. These variations in difficulty suggest that researchers designing an experiment seeking to evaluate map-symbolization methods should be sure to include these types of difficult questions in their testing.

If mapmakers know the purpose of a map, they should select a classification method well suited to the specific questions readers will have of the map. When working in a computer mapping environment, they can easily test multiple classifications to see how sensitive map patterns are to changes in classification (an approach long suggested by cartographers; e.g., Monmonier 1991). When possible, using matched legends aids map comparison with an impressive 28-percent improvement in accuracy in the reported results. Not all map series will lend themselves to this strategy, such as comparison of percentages to rates per 100,000. Use of color to allow many classes can also assist use of a common set of classes for a series with a wide range in data. The many classes that color permits improve the number of classes seen on individual maps that may span only a portion of the overall data range in the series.

In an era when maps are made from large databases with software that allows queries of individual polygons and iterative changes in classifications, it seems that facilitating map comparison is now more important than optimizing classification for a single map. Quantiles seem to be one of the best methods for facilitating comparison as well as aiding general map-reading. The rational advantages of careful optimization processes seem to offer no benefit over the simpler quantile method for the general map-reading tasks tested in the reported experiment.

Acknowledgments

We appreciate the careful and inspired work of Penn State graduate research assistants Cory Eicher and Erik Steiner on this research. We thank Robert Cromley for his assistance in providing code for the minimum boundary error algorithm. We would also like to acknowledge Alan MacEachren's critiques of the work. The research was supported by the National Center for Health Statistics/CDC, Office of Research and Methodology, Project #98-203. In addition, a portion of this material is based upon work supported by the National Science Foundation under Grant No. 9983451.

References

ArcView. Version 3.2. ESRI, Redlands, CA.

Armstrong, R. S. 1969. Standardized class interval and rate computation in statistical maps of mortality. *Annals of the Association of American Geographers* 59 (2): 382–90.

Becker, N. 1994. Cancer mapping: Why not use absolute scales? *European Journal of Cancer* 30A (5): 699–706.

Bertin, J. 1983. *Semiology of graphics: Diagrams, networks, maps*. Madison: University of Wisconsin Press.

Bregt, A. K., and M. C. S. Wopereis. 1990. Comparison of complexity measures for choropleth maps. *The Cartographic Journal* 27 (2): 85–91.

Brewer, C. A. 1994. Color use guidelines for mapping and visualization. In *Visualization in modern cartography*, ed. A. M. MacEachren and D. R. F. Taylor, 123–47. Tarrytown, NY: Elsevier Science.

———. 1996. Guidelines for selecting colors for diverging schemes on maps. *The Cartographic Journal* 33 (2): 79–86.

———. 1997a. Evaluation of a model for predicting simultaneous contrast on color maps. *The Professional Geographer* 49 (3): 280–94.
———. 1997b. Spectral schemes: Controversial color use on maps. *Cartography and Geographic Information Systems* 24 (4): 203–20.
———. 2001. Reflections on mapping Census 2000. *Cartography and Geographic Information Science* 28 (4): 213–35.
Brewer, C. A., A. M. MacEachren, L. W. Pickle, and D. J. Herrmann. 1997. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers* 87 (3): 411–38.
Brewer, C. A., and T. A. Suchan. 2001. *Mapping Census 2000: The geography of U.S. diversity*. Washington, DC: U.S. Census Bureau and Government Printing Office.
Carr, D. B., A. R. Olsen, and D. White. 1992. Hexagon mosaic maps for display of univariate and bivariate geographical data. *Cartography and Geographic Information Systems* 19 (4): 228–36, 271.
Carr, D. B., and L. W. Pickle. 1993. Plot production issues and details. *Statistical Computing and Statistical Graphics Newsletter* August:16–20.
Carstensen, L. W. 1986. Bivariate choropleth mapping: The effects of axis scaling. *The American Cartographer* 13 (1): 27–42.
Chang, K-T. 1978. Visual aspects of class intervals in choropleth mapping. *The Cartographic Journal* 15 (1): 42–48.
Coulson, M. R. C. 1987. In the matter of class intervals for choropleth maps: With particular reference to the work of George Jenks. *Cartographica* 24 (2): 16–39.
Cox, D. R. 1970. *Analysis of binary data*. London: Methuen & Co.
Cromley, E. K., and R. G. Cromley. 1996. An analysis of alternative classification schemes for medical atlas mapping. *European Journal of Cancer* 32A (9): 1551–59.
Cromley, R. G. 1995. Classed versus unclassed choropleth maps: A question of how many classes. *Cartographica* 32 (4): 15–27.
———. 1996. A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographic Information Science* 10 (4): 405–24.
Cromley, R. G., and R. D. Mrozinski. 1999. The classification of ordinal data for choropleth mapping. *The Cartographic Journal* 36 (2): 101–9.
Dent, B. D. 1999. *Cartography: Thematic map design*. 5th ed. Dubuque, IA: WCB/McGraw-Hill.
Dixon, O. M. 1972. Methods and progress in choropleth mapping of population density. *The Cartographic Journal* 9 (1): 19–29.
Dobson, M. W. 1973. Choropleth maps without class intervals: A comment. *Geographical Analysis* 5:358–60.
———. 1980. Perception of continuously shaded maps. *Annals of the Association of American Geographers* 70 (1): 106–7.
Egbert, S. L., and T. A. Slocum. 1992, EXPLOREMAP: An exploration system for choropleth maps. *Annals of the Association of American Geographers* 82 (2): 275–88.
Evans, I. A. 1977. The selection of class intervals. *Institute of British Geographers Transactions, New Series* 2 (1): 98–124.
Fisher, W. D. 1958. On grouping for maximum homogeneity. *Journal of the American Statistical Association* 53:789–98.
Gale, N., and W. C. Halperin. 1984. A case study for better graphics: The unclassed choropleth map. *The American Statistician* 36 (4): 330–36.
Gilmartin, P., and E. Shelton. 1989. Choropleth maps on high-resolution CRTs: The effect of number of classes and hue on communication. *Cartographica* 26 (2): 40–52.
Groop, R. E., and P. Smith. 1982. A dot-matrix method of portraying continuous statistical surfaces. *The American Cartographer* 9 (2): 123–30.
Jenks, G. F. 1963. Generalization in statistical mapping. *Annals of the Association of American Geographers* 53 (1): 15–26.
———. 1977. *Optimal data classification for choropleth maps*. Department of Geography Occasional Paper no. 2. Lawrence: University of Kansas.
Jenks, G. F., and F. C. Caspall. 1971. Error on choropleth maps: Definition, measurement, and reduction. *Annals of the Association of American Geographers* 61 (2): 217–44.
Jenks, G. F., and M. R. C. Coulson. 1963. Class intervals for statistical maps. *International Yearbook of Cartography* 3:119–34.
Kennedy, S. 1994. Unclassed choropleth maps revisited: Some guidelines for construction of unclassed and classed choropleth maps. *Cartographica* 31 (1): 16–25.
Kraak, M-J., and A. MacEachren. 1999. Visualization for exploration of spatial data. *International Journal of Geographical Information Science* 13 (4): 285–88.
Lavin, S., and J. C. Archer. 1984. Computer-produced unclassed bivariate choropleth maps. *The American Cartographer* 11 (1): 49–57.
Lewandowsky, S., and J. T. Behrens. 1995. Perception of clusters in mortality maps: Representing magnitude and statistical reliability. In *Cognitive aspects of statistical mapping*, ed. L. W. Pickle and D. J. Herrmann, 107–32. NCHS Working Paper Series, no. 18. Hyattsville, MD: National Center for Health Statistics.
Lloyd, R. E., and T. R. Steinke. 1976. The decision-making process for judging the similarity of choropleth maps. *The American Cartographer* 3 (2): 177–84.
Lloyd, R. E., and T. R. Steinke. 1977. Visual and statistical comparison of choropleth maps. *Annals of the Association of American Geographers* 67 (3): 429–36.
MacDougall, E. B. 1992. Exploratory analysis, dynamic statistical visualization, and geographic information systems. *Cartography and Geographic Information Systems* 19 (4): 237–46.
MacEachren, A. M. 1982. The role of complexity and symbolization method in thematic map effectiveness. *Annals of the Association of American Geographers* 72 (4): 495–513.
———. 1985. Accuracy of thematic maps/Implications of choropleth symbolization. *Cartographica* 22 (1): 38–58.
———. 1994. *Some truth with maps: A primer on symbolization and design*. Washington, DC: Association of American Geographers.
———. 1995. *How maps work: Representation, visualization, and design*. New York: Guilford.
MacEachren, A. M., C. A. Brewer, and L. W. Pickle. 1998. Visualizing georeferenced data: Representing reliability of health statistics. *Environment & Planning A* 30 (9): 1547–61.
MacEachren, A. M., and D. DiBiase. 1991. Animated maps of aggregate data: Conceptual and practical problems. *Cartography and Geographic Information Systems* 18 (4): 221–29.
Mackay, J. R. 1955. An analysis of isopleth and choropleth class intervals. *Economic Geography* 31 (1): 71–81.
Mak, K., and M. R. C. Coulson. 1991. Map-user response to computer-generated choropleth maps: Comparative experiments in classification and symbolization. *The American Cartographer* 18 (2): 109–24.
McGranaghan, M. 1989. Ordering choropleth map symbols: The effect of background. *The American Cartographer* 16 (4): 279–85.
Mersey, J. E. 1990. Colour and thematic map design: The role of

colour scheme and map complexity in choropleth map communication. *Cartographica* 27 (3): 1–157.
Monmonier, M. S. 1972. Contiguity-based class-interval selection: A method of simplifying patterns on statistical maps. *Geographical Review* 62:203–28.
———. 1973. Analogs between class-interval selection and location-allocation models. *The Canadian Cartographer* 10 (2): 123–32.
———. 1974. Measures of pattern complexity for choropleth maps. *The American Cartographer* 1 (2): 159–69.
———. 1975. Class intervals to enhance the visual correlation of choroplethic maps. *The Canadian Cartographer* 12 (2): 161–78.
———. 1982. Flat laxity, optimization, and rounding in the selection of class intervals. *Cartographica* 19 (1): 16–26.
———. 1991. Ethics and map design: Six strategies for confronting the traditional one-map solution. *Cartographic Perspectives* 10:3–8.
———. 1992. Authoring graphic scripts: Experiences and principles. *Cartography and Geographic Information Systems* 19 (4): 247–60, 272.
———. 1994. Minimum-change categories for dynamic temporal choropleth maps. *Journal of the Pennsylvania Academy of Science* 68 (1): 42–47.
Muller, J-C. 1976. Number of classes and choropleth pattern characteristics. *The American Cartographer* 3 (2): 169–76.
———. 1979. Perception of continuously shaded maps. *Annals of the Association of American Geographers* 69 (2): 240–49.
Muller, J-C., and J. L. Honsaker. 1978. Choropleth map production by facsimile. *The Cartographic Journal* 15 (1): 14–19.
Olson, J. 1972a. Class-interval systems on maps of observed correlated distributions. *The Canadian Cartographer* 9 (2): 122–32.
———. 1972b. The effects of class-interval systems on choropleth map correlation. *The Canadian Cartographer* 9 (1): 44–49.
———. 1975a. Autocorrelation and visual map complexity. *Annals of the Association of American Geographers* 65 (2): 189–204.
———. 1975b. Spectrally encoded two-variable maps. *Annals of the Association of American Geographers* 71 (2): 259–76.
Olson, J., and C. A. Brewer. 1997. An evaluation of color selections to accommodate map users with color vision impairments. *Annals of the Association of American Geographers* 87 (1): 103–34.
Paslawski, J. 1983. Natural legend design for thematic maps. *The Cartographic Journal* 20 (1): 36–39.
———. 1984. In search of a general method of class selection for choropleth maps. *International Yearbook of Cartography* 34:159–69.
Peterson, M. P. 1979. An evaluation of unclassed crossed-line choropleth mapping. *The American Cartographer* 6 (1): 21–38.
Pickle, L. W., M. Mungiole, G. K. Jones, and A. A. White. 1996. *Atlas of United States mortality*. Hyattsville, MD: National Center for Health Statistics.
Robinson, A. H., J. L. Morrison, P. C. Muehrcke, A. J. Kimerling, and S. C. Guptill. 1995. *Elements of cartography*. 6th ed. New York: John Wiley & Sons.
SAS PROC LOGISTIC. SAS Institute, Cary, NC.
Scripter, M. W. 1970. Nested-means map classes for statistical maps. *Annals of the Association of American Geographers* 60 (2): 385–93.
Slocum, T. A. 1999. *Thematic cartography and visualization*. Upper Saddle River, NJ: Prentice-Hall.
Slocum, T. A., and S. L. Egbert. 1993. Knowledge acquisition from choropleth maps. *Cartography and Geographic Information Systems* 20 (2): 83–95.
Slocum, T. A., S. H. Robeson, and S. L. Egbert. 1990. Traditional versus sequenced choropleth maps: An experimental investigation. *Cartographica* 27 (1): 67–88.
Smith, R. M. 1986. Comparing traditional methods for selecting class intervals on choropleth maps. *The Professional Geographer* 38 (1): 62–67.
Stegna, L., and F. Csillag. 1987. Statistical determination of class intervals for maps. *The Cartographic Journal* 24 (2): 142–46.
Steinke, T. R., and R. E. Lloyd. 1981. Cognitive integration of objective choropleth map attribute information. *Cartographica* 18 (1): 13–23.
Steinke, T. R., and R. E. Lloyd. 1983. Judging the similarity of choropleth map images. *Cartographica* 20 (4): 35–42.
Tobler, W. R. 1973. Choropleth maps without class intervals? *Geographical Analysis* 5:262–64.
Walter, S. D., and S. E. Birnie. 1991. Mapping mortality and morbidity patterns: An international comparison. *International Journal of Epidemiology* 20 (3): 678–89.

Correspondence: Department of Geography, The Pennsylvania State University, University Park, PA 16802-5011, e-mail: cbrewer@psu.edu (Brewer); Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD 20892-8317, e-mail: picklel@mail.nih.gov (Pickle).