PARALLEL COORDINATE PLOTS FOR REPRESENTING DISTRIBUTION SUMMARIES IN MAP LEGENDS

Daniel B. Carr
Center for Computational Statistics, 4A7
George Mason University
Fairfax, VA 22030
U.S.A.
Phone: (703) 993-1671
Fax: (703) 993-1700
Email: dcarr@voxel.galaxy.gmu.edu

Anthony R. Olsen
U.S. EPA Environmental Research laboratory
200 S.W. 35th St.
Corvallis, Oregon 97333
U.S.A.
Phone: (503) 754-4790
Fax: (503) 754-4716
Email: tolsen@heart.cor.epa.gov

Abstract

This paper addresses the graphical representation of distributional summaries. The graphical design goal is to produce small summary plots that are suitable as a map legends for Choropleth maps. The paper proposes two variations of parallel coordinate plots for representing cumulative distributions. The first variation shows the cumulative distribution and represents the distribution density using color. The second variation shows class colors and adds distribution detail within class boundaries. This second form is suitable as a map legend. The paper provide examples of the plot used as a legend in Choropleth maps of acid deposition and colon cancer mortality.

1 Introduction

Distributional summaries, whether specifically part of map legends or not, facilitate interpretation of attribute values represented in maps. For example, analysts studying a cancer death rate map for U.S. counties may want to know if the rate for a specific county is above the 95th population weighted percentile for the nation. Distributional summaries provide a basis for answering such questions.

The graphics for representing distributional summaries take many forms: density plots, histograms, quantile plots, empirical cdf plots, and even normal qqplots. Each graphic type has various limitations and strengths for purposes of relating specific values to probabilities and vice versa. This paper describes two new plots for representing cumulative probability (p) and quantile (q) pairs.

Traditional quantile and cdf plots represent pq pairs using Cartesian coordinates. Given the same pq pairs, the Cartesian plots are basically equivalent. The cdf plot puts probabilities on the vertical axis and the quantile plot puts probabilities on the horizontal axis. The proposed parallel coordinate approach uses two vertical axes. We call the plots pq plots or qp plots depending on the left-to-right order of the two axes.

Parallel coordinate plots provide an alternate approach to representing number pairs. The basic idea is to construct parallel axes and to connect the coordinates of point pairs using a straight lines. Two key papers describe the geometry and interpretation of parallel coordinates plots.[1,2] The current application is particularly simple. Since the cdf is a function, lines for distinct points pairs never cross between the axes. Lines can intersect on the probability axis. That is, for discrete distributions a probability may connect to an interval on the quantile axis. However, the expected practice is to show connecting lines just for selected jump points. Lines appearing to intersect on the quantile axis can only be low resolution artifacts. For table look-up purposes following straight lines is easy. The absence of crossing lines makes the task even easier.

This paper calls attention to two variations, the pq density plot and the pq piecewise linear plot. The pq density plot represents a surface created by interpolating densities along the pq lines between parallel axes. The second variation, the pq piecewise linear plot uses the region between the parallel axes to show class colors. This distributional summary retains substantial detail and is suitable for use as a map legend.

2 CDF Plot and Legend Limitations

Cdf plots similar to that in Figure 1 (the grid is typically omitted) provide distributional summaries and appear in reports by EPA and other government agencies. While commonly used, such plots prove awkward in regard to several tasks. The awkward tasks include looking up value pairs and, in the map legend context, showing color links to Choropleth map classes.

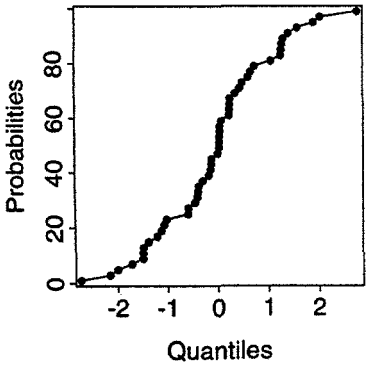

Figure 1: A CDF Plot

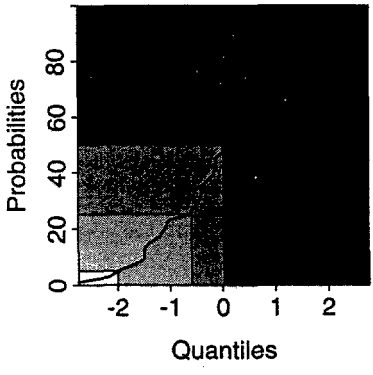

Figure 2: A Legend Attempt

Consider the table look-up process in Cartesian coordinate plots. The visual path from quantile to curve to probability (or vice versa) involves a right angle turn. The visual path length differs markedly from small to large quantiles so the treatment of quantiles is not uniform. Standard horizontal and vertical grid lines based on pretty axis values do not typically intersect on the cdf curve and thus do not directly support the reading of specific pq pairs.

If the plot includes right-angle reference lines from the quantile scale to the cdf curve to the probability scale, interpolation may still be required to "read" the values of one or both members of each reference pair. Adding reference lines and labeling their endpoints is a possibility, but linear scales often leave little space for labeling, especially along the horizontal axis. If skewness provides labeling space for one tail of the distribution, it robs space from the rest of the distribution. Common cdf plots show no reference pairs.

For typical cdf plots, readers must expend mental energy to obtain verbally expressed pq (or qp) pairs. In an application setting, such energy could be put to better use in memorizing a few values for later reference or in comparing values to other distributions. The Cartesian coordinate representation may seem to be a good storage device but is less than ideal for quick reading value pairs. We conjecture that few people read more than one or two pairs from a typical cdf plot.

Data analysts often find Choropleth maps more informative if they include distributional summaries of the attributes represented in the maps. However, the inclusion of cdf plots is awkward for two reasons. First, the cdf plot takes up a large area relative to the linear resolution of the two axes. Second, the addition of class colors to a cdf plot is a design challenge. Figure 2 provides a gray-scale caricature of using the plot region to show class definitions. The disportionately large areas for large quantiles are not acceptable. A second choice is to add colored rectangles along the probability (or quantile) axis when the probabilities (or quantiles) define the classes. The smallest of these rectangles has to be of sufficient area so that readers can easily perceive its color. Adding color rectangles along an axis takes up more space as well as complicating the placement of tics and tic labels. The Cartesian coordinate approach is less than optimal for use as a legend.

Typical cartographic legends show the class colors in rectangles. Map makers label these rectangles with quantile (value) bounds or probability (percent) bounds but not typically both.[3] Some maps show both quantile and percent legends.[4] By looking from legend to legend and focusing on corresponding class boundaries one can figure out a few pq pairs. Most of the distributional information is lost. Cartographic legends typically provide little distributional information.

3 Estimation Issues

Before describing pq plots in detail, a few comments on the estimation of probabilities seem appropriate. Computing probabilities from samples is an essential task. For a simple random sample two estimation approaches are common, the empirical cdf approach and the distribution-of-order-statistics approach.

The empirical cdf for a sample of size n is

$$P(x) = \quad i/n \qquad x_{(i)} \le x < x_{(i+1)} \text{ for } i=0,\ldots,n \tag{1}$$

where $x_{(i)}$ are order statistics, $x_{(0)}=-\infty$ and $x_{(n+1)}=\infty$. The definition is troublesome in the tails. That is, the estimated probability for future observations more extreme than the sample extrema is zero. Such probability estimates for extreme values are biased low and hence counter-intuitive. While theory shows that the bias approaches zero as the sample size approaches infinity most people work with finite samples. Recognizing the possibility of more extreme values seems reasonable.

Order statistics results [5,6,7] suggest a more plausible expression:

$$P(x) = \quad (i-a)/(i+1-2*a) \qquad \text{for } x \cong x_{(i)},\ i=1,\ldots,n \text{ and } 0 \le a \le .5 \tag{2}$$

A straight-forward approach obtains probability estimates for quantiles between the order statistics by linear interpolation and estimates beyond the sample extrema by extending the probabilities at the sample extrema.[8] The estimates for the current graphics use this approach with a=.5. The software producing the graphics will require revision when linear interpolation produces poor estimates.

Even with a distribution-of-order-statistics approach, probability estimates near extrema are uncomfortably close to pure guesses given the amount of scrutiny they may receive. The proposed graphic design allows users to finesse the issue by specifying quantiles or probability limits for the plot. The plot labeling can then list the sample extrema or user-imposed limits without attaching the corresponding probability estimates.

4 The PQ Density Plot

The pq density plot (pqd plot) applies to distributions with density functions. The plot construction follows from a few simple observations. First, we can compute a density along each axis. Then we can interpolate the density along pq lines between the axes. The construction of density estimates for the quantile axis is a well-studied problem.[9] We can choose from many methods. The probability integral transform states that the density is uniform on the probability axis. Since we assume a density function for the quantile axis, the cdf has no jump points. Consequently we can pick any point between the p and q axes and find the pq line that goes through the point. Doing this for a rectangular lattice of points between the axes and interpolating densities between line endpoints yields a density image.

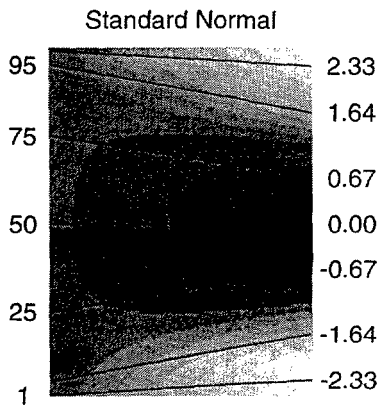

Figure 3a

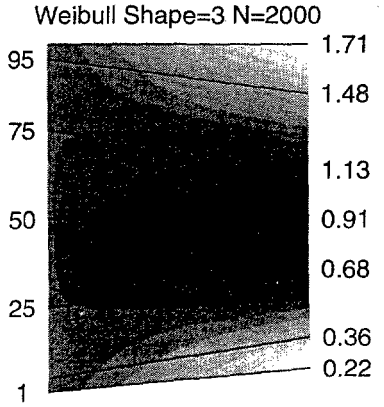

Figure 3b

Figure 3a is a pqd plot for a truncated standard normal distribution. The figure shows a few standard pq lines and the labeling for the p axis shows percents rather than probabilities. In this gray-level version the gray level on the p axis represents the average density on the q axis. Lighter shades of gray indicate lower density and dark shades indicate higher density. A pair of qp lines (right to left) starting in a light gray region must converge (or compress the area between the lines) to achieve the uniform density value. A pair of qp lines starting in a dark gray region must diverge (or expand the area between lines) to achieve the uniform density value. This type of plot can help student intuition. Figure 3a is special case because the distribution is theoretical and symmetric.

Figure 3b uses function estimates based on a sample of 2000 points from a Weibull distribution. This distribution is not symmetric. Figure 3b illustrates a particular scaling choice for the axes. The choice forces the median line to be horizontal. The user specified quantile bound furthest from the median provides a second point and the two points determine the linear graphic scale. This scale leaves the q axis empty beyond the short tail and Figure 1b shows the empty region in white. The pq density plot applies to both theoretical and empirical distributions.

Annual 1985-1987 Sulfate Deposition

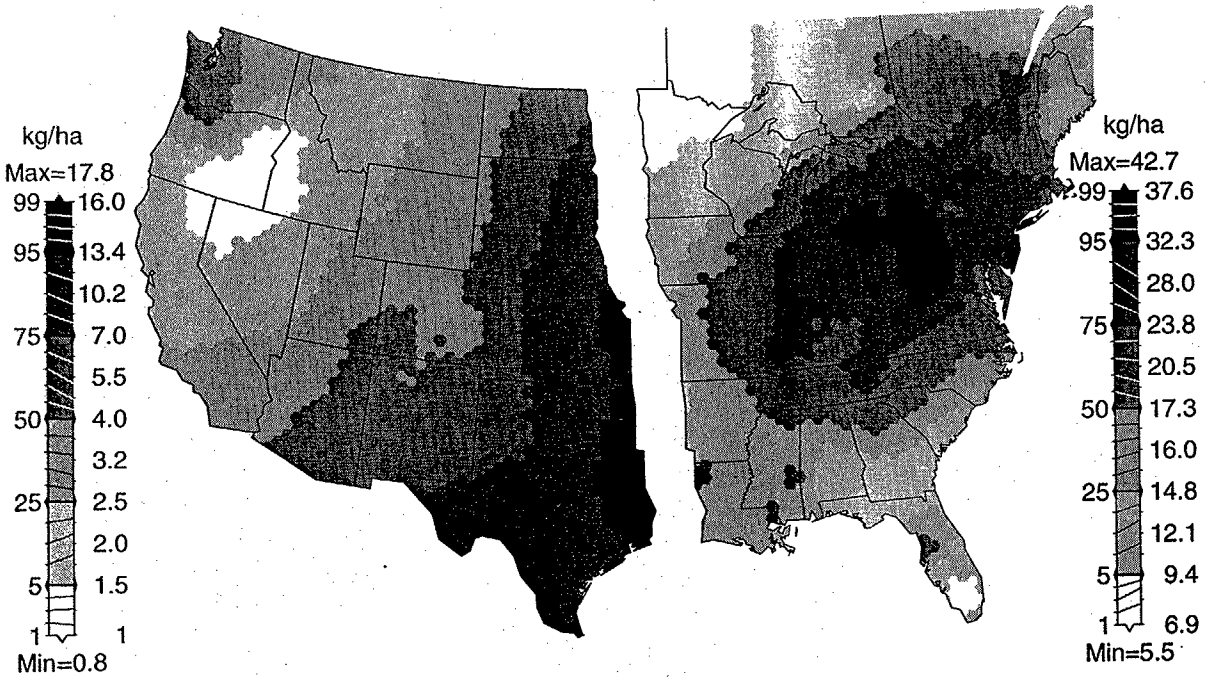

Figure 4: Legend Distribution for Area

White Male Colon Cancer
Health Service Areas For 1980-1989
Age-Adjusted Rates Per 100,000

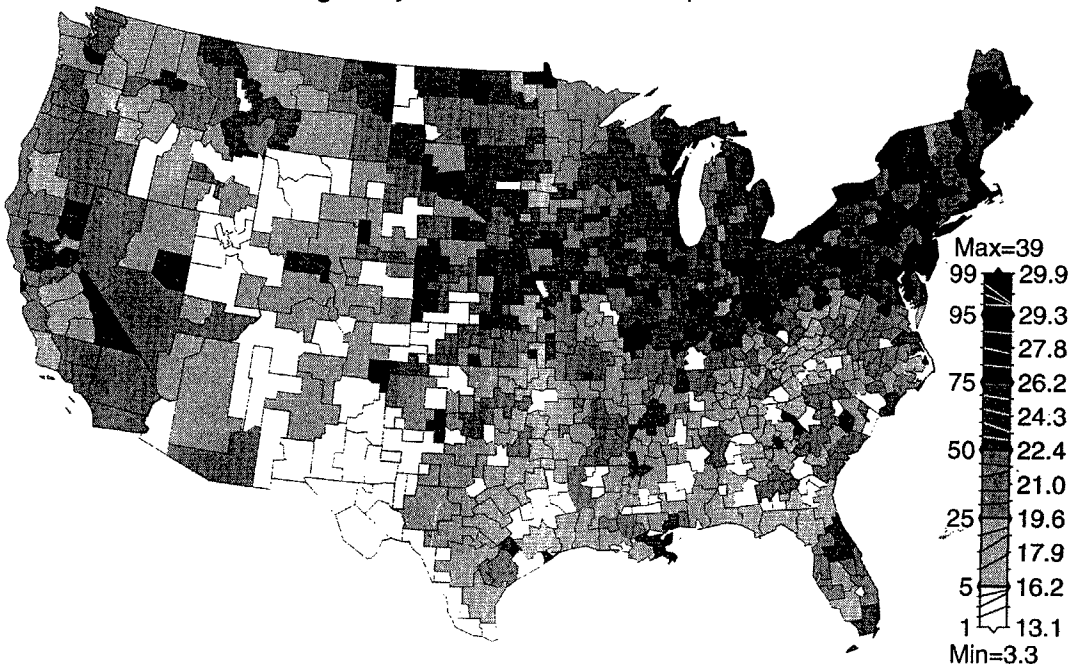

Figure 5: Legend Distribution Relative to Population Size

The construction of the pqd plot brings together concepts of cumulative probabilities, quantiles, the probability integral transformation, order statistics, densities, interpolation, image construction and surface representation. A perspective view from the quantile side provides more geometric intuition about the data density than can be obtained by color alone. A fully rendered color surface with highlights and reference lines would look even better.

While pqd plots are of interest as a distributional summaries, they are not well-suited as legends for Choropleth maps. They do not represent the class colors and they have some labeling problems. Inspection of the plots reveals omission of the 5 and 99 percent labels. The software drops the labels due to the lack of plotting space in plots of the this size. The problem is not just a coding artifact. Percents such as 1 and 5 are going to be close on any small plot with a linear percent scale. This labeling problem motivates the next variation, the pq piecewise linear plot.

5 The PQ Piecewise Linear Plot

The pq piecewise linear plot (pqp plot) is a generalization of previous legends.[10] The previous legends had a linear p axis and represented key pq pairs using horizontal lines. The key pq pairs the formed boundaries between the Choropleth classes. The rectangles between the axes and the pq lines showed the class colors. That approach worked for showing class colors. However, the implicit nonlinear q axis made it impossible to determine the values for additional pq pairs.

The new plot appears as map legends in Figures 4 and 5. The fundamental change is that the two axes are now piecewise linear. The black triangles and thick horizontal black lines mark the division between the linear sections. Note that the p axis is much larger in regions above 95 percent and below 5 percent than it would be on a linear scale. This facilitates showing more detail in the regions that are often of most interest. The regular spacing of q axis tics and tic labels determines the p axis scale in the interior of the plot.

The plot design encourages reading from percents on the left to percentiles (quantiles) on the right. The addition of tics and references lines allows readers to obtain additional pq pairs by interpolation. The interpolation always occurs on a linear scale between class boundaries. The reference lines have 5 percent increments in the center and 1 percent increments near the tails. In addition to being interpolation aids, these reference lines help show skewness and distributional anomalies.

Triangles at top and bottom of the plots point to extrema. The user selects either quantile limits for the quantile axis or the probability limits for the probability axis and this determines the corresponding limits for the other axis. Either approach can exclude the sample extrema, increase resolution for values represented, and finesse uncertainties in calculating probabilities for extrema.

Researchers who collect the data represented on maps are inclined to find most map legends impoverished. Those who study maps often want more distributional detail. The pqp plot includes more detail without taking up much space. The plot provides a compromise between simple legends and extensive pq tables.

6 Conversion From Color To Gray Scale

Most researchers will produce full color maps on their monitors if not in print. The originals for Figures 4 and 5 were in color and the gray-scale versions do not show the pqp plots to full advantage. The conversion to gray-scale introduced some complications that warrant comment.

The original for Figure 4 had additional class boundaries at 90 percent and 10 percent. The variety provided by different hues and saturation levels supports the use of more classes. We dropped the two

class boundaries because more than six gray-levels seemed too much. Even with six classes, matching the map grays to legend grays is complicated by surround induced changes in perceived gray level. Guidance about surround-influenced matching problems applies primarily to the more general color context.[11] With gray levels not much can be done beyond limiting the number of classes and increasing the contrast between classes.

Something else is lost in Figure 4. The original map used value-ordered shades of blue for the Western U.S. and a value-ordered sequence from yellow to orange to brown for Eastern North America. We split the map because the distribution of values differed radically for the two parts of the U.S. Showing both regions in gray approximates the value ordering but turns the striking differences between West and East into a subtle legend distinction.

The original for Figure 5 used a double-ended scale with blues for death rates below the median and reds for death rates above the median. The two extreme classes were saturated colors and the two middle classes were close to light gray. While the blue to gray to red scale is not bad, recent research suggests that two other double-ended scales may be preferable. One scale is a blue to gray to orange scale. The other is a green to gray to purple scale. People with two major types of color blindness can still used these scales.[12] For Figure 5, we dropped the double-ended scale since it does not make sense when rendered in gray.

7 Interpretation

Three distributions are often relevant to the display of a Choropleth map.[13] These are the distributions of the map's attribute values relative to the number of regions, the area of regions, and the number of people in the regions. In some cases showing all three pqp plots on the same map warns about the discrepancy between the area-based visual impact and the appropriate interpretation. Figures 4 and 5 illustrate the different distributions.

Figure 4 shows sulfate deposition and the measurement units are kilograms per hectare. The percents on the left side of the legend refer to the percent of total area with values below the corresponding quantile. In an equal area projection, regular hexagons all have the same area. Thus the area distribution is equivalent to the distribution for the number of hexagons. In other words the pqp plot in Figure 4 simultaneously represents two equivalent distributions.

As for a brief comment on interpretation, the Canadian position has been that values above 20 kilograms per hectare are cause for concern. The map shows values in Eastern North American above this criterion. However the area is difficult to assess from the map because 20 kilograms per hectare in not a class boundary. However, the legend allows are reader to make a reasonable estimate of the area for this or other criteria that may be proposed.

Figure 5 concerns cancer death rates for health service areas. The unit of measure is deaths per 100,000 white males. The legend provides the percent of white males living in health service areas with rates below the corresponding quantiles.

The regions called health service areas may not be familiar. The U.S. National Center for Health Statistics uses health service areas in many of their mortality maps. The health service areas are counties or aggregates of counties. The goal in aggregating counties is to produce fewer regions with more homogeneous population sizes. Neighboring counties whose people receive health services from the same local institutions are subject to aggregation.

Figure 5 immediately shows a pattern of high colon cancer death rates in the North East. This pattern has

been apparent for over three decades. Plotting the residuals from a spatial smooth of the cancer rates brings out additional patterns in the data.[14] The residuals call attention to local anomalies in the cancer rate surface and are usefully for generating hypothesis. Two local hot spots (high residuals) have a high percentage of people with Czechoslovakian background. An epidemiological study of one hot spot points to dietary patterns as a cause of increased mortality.[15]

8 Evaluation and Extensions

With an already burdensome variety of methodological alternatives, new methods proposed for use should be demonstrably superior to existing alternatives in some significant domain. In this paper we suggest that the pqd and pqp plots have sufficient merit to warrant serious consideration. In particular the pqp plot is easy to understand and well-suited for serving as a map legend and a distributional summary.

The pqp plot has not yet been subjected to cognitive studies. In may be that slight variations are preferable. For example, separating the piecewise sections may have merit even though the approach takes up more space and may require labeling. Additional labeling can clarify the treatment of ties at class boundaries. Careful color selection may reduce color matching errors. Cognitive testing is an important next step.

The ability of plotting methods to extend to other situations is also an important consideration. Figure 4 raises the question of comparing two distributions, the distributions for the Western U.S. and Eastern North America. The qq plot provides a preferred way of comparing two distributions. In view of the above discussion, a parallel coordinates qq plot seems an obvious next step and a prime candidate to assist in the comparison of attributes represented in Choropleth maps.

Figures 4 and 5 show a statistician's bias in that conventional percents define the class boundaries.. However, there is no problem in defining boundaries based on quantiles. This suggests the obvious variation, qp plots. In qp plots the boundary defining quantiles appear on the left and will often be convenient integer values.

Another extension involves the representation cdf confidence bounds. The confidence curves for traditional cdf plots are deceptive. Humans do not judge distances between curves in the correct vertical direction but rather assess distances in a direction roughly normal to the curves.[16] While a thin pqp plot may not be able to represent all the confidence bound estimates shown in a traditional cdf plot, we anticipate that pqp plots can represent selected bounds effectively.

Readers interested reproducing the figures in this paper or in adapting methods to their own applications can obtain Splus functions and example script files. Use anonymous ftp to galaxy.gmu.edu and look in directory /pub/submissions/pq.

Acknowledgments

Research related to this article by EPA under cooperative agreement No. CR8280820-01-0. The article has not been subjected to the review of the EPA and thus does not necessarily reflect the view of the agency and no official endorsement should be inferred.

References

[1] Inselberg, A. 1985. The Plane With Parallel Coordinates, " *The Visual Computer*, 1, pp. 69-91.

[2] Wegman, E. J. 1990. "Hyperdimensional Data Analysis Using Parallel Coordinates", *Journal of the American Statistical Association*, Vol. 85. No 411. pp. 664-675.

[3] Dent, D. B. 1990. *Cartography, Thematic Map Design.* Wm. C. Brown Publishers. Dubuque, Iowa.

[4] Goldman, B A. 1991. *The Truth About Where You Live,* Times Books, Random House Inc. New York.

[5] Blom, G. 1958. *Statistical Estimates and Transformed Beta-Variables.* John Wiley and Sons. New York.

[6] David, H. A. 1970. *Order Statistics.* John Wiley and Sons. New York.

[7] Hoaglin, D. C 1983. "Letter Values: A Set of Selected Order Statistics" in *Understanding Robust and Exploratory Data Analysis,* Editors: Hoaglin, Mosteller and Tukey, John Wiley and Sons, Inc. New York. pp 33-57.

[8] Chambers, J. M, W. S. Cleveland, B. Kleiner, P. A. Tukey. 1983. *Graphical Methods for Data Analysis,* Wadsworth and Brooks/Cole, Pacific Grove, California.

[9] Scott, D. W. 1992. *Multivariate Density Estimation, Theory, Practice and Visualization.* John Wiley and Sons, Inc. New York.

[10] 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,* Vol. 19, No. 4, pp. 228-236,271.

[11] Brewer, C. A. 1991. *Prediction of Surround-Induced Changes in Map Color Appearance,* Doctoral Dissertation, Department of Geography, Michigan State University.

[12] Brewer, C.A , A. M MacEachren, and L. W. Pickle. 1995. Evaluation of Map Color Schemes of the NCHS Mortality Atlas, *Proceedings of the International Symposium on Computer Mapping in Epidemiology and Environmental Health,* Tampa FL, In Press.

[13] Carr, D. B, 1993. Constructing Legends for Classed Choropleth Maps." Statistical Computing & Statistical Graphics Newsletter, Vol. 1. No 1, pp. 15-19.

[14] Carr, D. B. 1994. Color Perception. the Importance of Gray and Residuals on a Choropleth Map. *Statistical Computing & Graphics,* Vol 5. No. 1, pp. 17-20.

[15] Pickle, L.W., T. J. Mason, N. Howard, R. Hoover, and J. R. Fraumeni, Jr. 1987. *Atlas of U..S. Cancer Mortality Among Whites: 1950-1980.* Washington, D.C.: USGPO, DHHS Publ. No. (NIH) 90-1582

[16] Cleveland, W. S. 1985. *The Elements of Graphing Data.* Wadsworth. Monterey, California.