Analysis and Visualization

After the extracted features have been identified and classified catalogued, they are examined visually and then analyzed. Features may be marked at this stage for use in the learning phase. Visualization tools include a Feature Dendogram Catalog (FDC), a Time Between Events Plot (TBEP), and a Feature Plot. The FDC Plot provides a graphical catalog of the features clustered according to the attributes used at the extraction stage. For each feature or cluster of features, the TBEP shows the time since the feature cluster was last observed. This plot is useful for detecting whether certain feature clusters are periodic. As its name implies, the Feature Plot shows the segment of the time series that contains the feature and its context.

Learning

At the end of the analysis and visualization stage, all identified features are classified into one of two groups: those that were flagged at the analysis and visualization stage and those that were not. At this stage, the FEaTureS code begins an optimization based on the ability of each feature identification tool to classify the features correctly. The optimization process is repeated for each of the available tools. This process may be viewed as a multivariate optimization in which the objective is to minimize the number of misclassified features. The optimization results will differ both quantitatively and qualitatively from one tool to another. From a summary of these results, the scientist may decide which "optimized" tool to use.

In Conclusion

With today's automated data collection systems, it is possible to assemble enormous data sets with comparative ease. The ability to acquire data is outstripping our capacity to analyze them. The result is that in many cases only a small fraction of the data is ever examined. Automated data monitoring systems such as those described here offer one way method for recovering the interesting features of large data sets before they are lost forever.

Acknowledgments

I thank my colleagues at Oak Ridge National Laboratory for allowing me to share this description of their work in progress.

Albert M. Liebetrau
Battelle Pacific Northwest Laboratories
AM_Liebetrau@pnl.gov

TOPICS IN SCIENTIFIC VISUALIZATION

Parallel Coordinate Variants Of CDF and Quantile Plots

by Daniel B. Carr and Anthony R. Olsen

Introduction

This article 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 straight lines. Two key papers, Inselberg (1985) and Wegman (1990), describe the geometry and interpretation of parallel coordinates plots. 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 article calls attention to two *pq* plot variations, the *pq* density plot and the *pq* piecewise linear plot. Carr and Olsen (1995) describe additional variations and provide construction details. The *pq* density plot represents the surface created by interpolating densities along the *pq* lines between parallel axes. When shown as color images (see Figures 1a and 1b) or as rendered surfaces, such colorful plots draw student interest. The second variation, the *pq* piecewise linear plot (see Figures 2a and 2b), is more of a visual table. This distributional summary retains substantial detail and is suitable for use as a map legend.

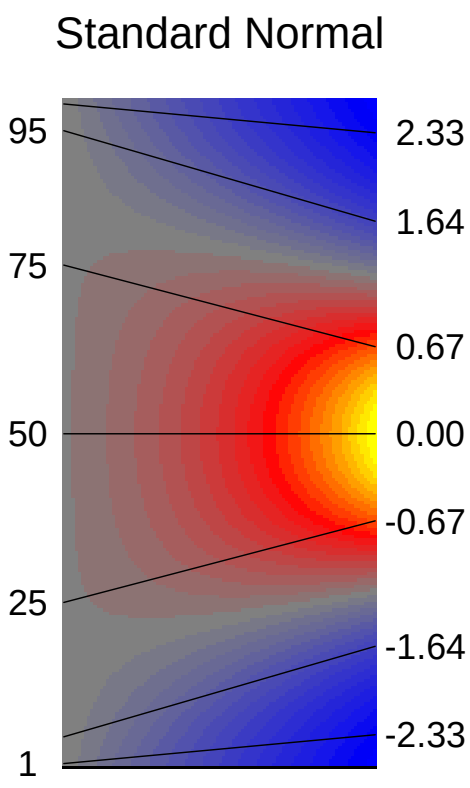

Figure 1a

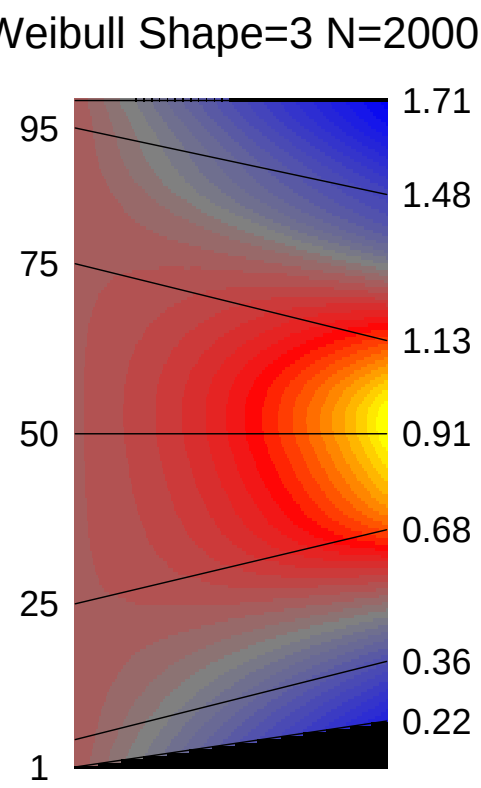

Figure 1b

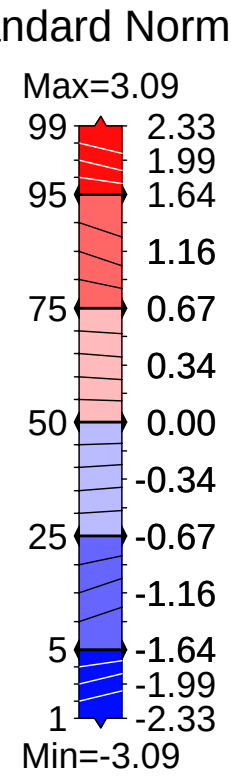

Figure 2a

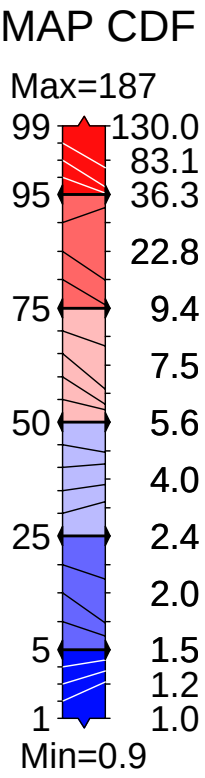

Figure 2b

CDF Plot and Legend Limitations

CDF plots similar to that in Figure 3 (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.

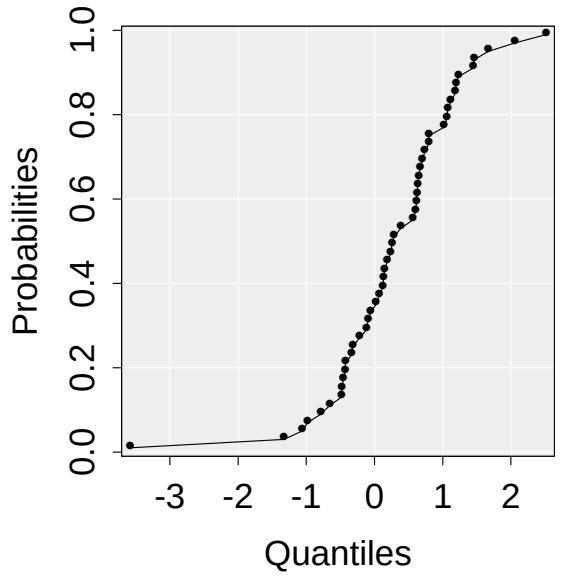

Figure 3: A CDF Plot

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 be 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 of 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 statistical summaries of the spatial phenomena. The *pq* pairs can represent a variety of summaries such as the population size in different classes, the map area in different classes, and the number of regions in different classes (see Carr 1993). Having chosen an appropriate summary, a statistician's first thought might be to add class colors to a cdf plot and use it as a legend. However, cdf plots are awkward for this task 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 4 uses gray levels in the plotting 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 the reader can easily perceive its color. Adding colored 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 percent (probability) bounds but not typically both (for example see Dent 1992). Goldman (1991) shows both quantile and percent legends. 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. Typical legends provide poor statistical summaries.

Estimation Issues

Before further describing the example *pq* plots, 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

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

where $x_{(i)}$ are order statistics, $x_{(0)}$=-∞ and $x_{(n+1)}$=∞. The definition is troublesome in the tails. That is, the estimated probability for future observations more ex-

treme 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.

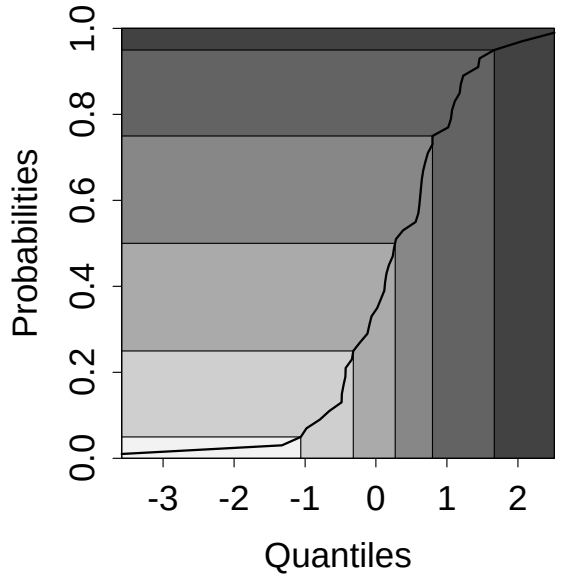

Figure 4: A Legend Attempt

Order statistics results (see Blom 1958, David 1972 and Hoaglin 1983) suggest a more plausible expression:

$$P(x) = (i - a)/(i + 1 - 2 * a) \quad \text{for } x = x_{(i)}$$

$$i = 1, ..., n \text{ and } 0 \leq a \leq .5$$

Chambers et al (1983) suggest obtaining 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. The estimates for the current graphics use this approach with a=.5. The software referenced by this article 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.

The PQ Density Plot

The *pq* density 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 (see Scott 1992). We can chose 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 1a is a *pq* density plot for a truncated standard normal distribution. The figure shows a few standard *pq* lines. The labeling for the *p* axis shows percents rather than probabilities. With minor exceptions the color assignment from blue to gray to red to yellow represents increasing densities with increasingly brighter colors (see Carr 1994). The color assignment fixes the number of color levels so that the "average density" on the *q* axis is gray. Correspondingly the whole *p* axis is gray. The blue regions on the *q* axis indicate below average density. The red and yellow regions indicate above average density. A pair of *qp* lines (right to left) starting in a blue region must converge (or compress the area between the lines) to achieve the uniform density value. A pair of *qp* lines starting in a red or yellow region must diverge (or expand the area between lines) to achieve the uniform density value. The plot can help student intuition. The color scale may not be the students' favorite and the opportunity to experiment with colors may lead to more than a passing glance at such plots.

The construction of Figure 1a uses theoretical density and quantile functions. In contrast Figure 1b uses function estimates based on a sample of 2000 points from a Weibull distribution. This distribution is not symmetric. Figure 1b 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 black. The color assignment in Figure 1b should be adjusted so that the *p* axis is gray. (The process of generating both images directly on the same page introduced a color reassignment puzzle that

we have not yet solved.) The *pq* density plot applies to both theoretical and empirical distributions.

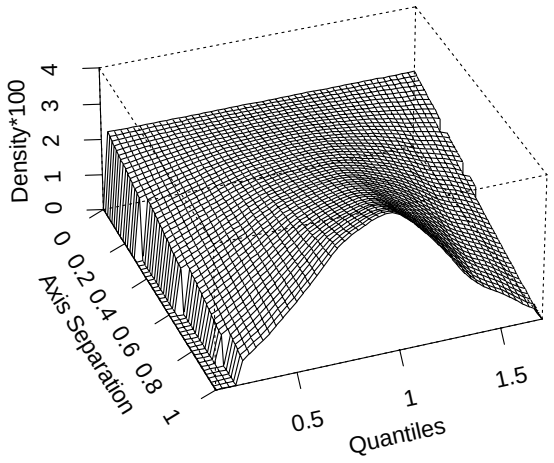

Figure 5: A Perspective View

The construction of the *pq* density plot brings together concepts of cumulative probabilities, quantiles, the probability integral transformation, order statistics, densities, interpolation, image construction and surface representation. Figure 5 shows a perspective view from the quantile side. This view provides more geometric intuition. The mesh would be better if it showed *pq* lines rather than a rectangular grid of lines. A fully rendered color surface with highlights and reference lines would look even better. The construction of different *pq* density views is an instructive exercise.

Inspection of the two plots reveals some omissions. The small plot size pretty well hides the dropped pixels along edges. The pixel problem can be fixed. The distressing omissions are the 5 and 99 percent labels. Due to the lack of plotting space, the software drops the labels. 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.

The PQ Piecewise Linear Plot

The *pq* piecewise linear plot is a generalization of the legends shown in Carr, Olsen and White (1992). The previous legends had a linear *p* axis and represented key *pq* pairs using horizontal lines. Their approach implied a nonlinear *q* axis and readers could not estimate values for other *pq* pairs. The new version shown in Figures 2a and 2b is a set of vertically juxtaposed linear *pq* plots. The black triangles and thick horizontal black lines mark the division between the linear plots. The composite *p* axis is not linear. For example the regions above 95 percent and below 5 percent are larger than they would be on a linear scale. This facilitates labeling and showing more detail in the regions that are often of most interest.

When the plot is a legend for a Choropleth map the key *pq* pairs are the boundaries between the Choropleth classes. The regions between the axes then show the class colors.

Triangles at the 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.

The plot design encourages reading from percents to percentiles (quantiles). The design provides lines for standard percents with 5 percent increments in the center and 1 percent increments near the tails. This generally corresponds to reader interest. For example, the 96^{th} percentile is usually of more interest than the 51^{th} percentile. For visibility the reference line colors need to contrast with the class colors. Figure 2a and 2b uses white lines for the extreme classes. This provides an additional cue about the special treatment of the extreme classes. Following the lines between the axes is easy.

To determine values in addition to the key *pq* pairs readers must interpolate quantiles using the probability-based reference lines. While interpolation is not trivial, the numerous quantile tics and tic labels provide bounds on the answer. The interpolation always occurs on a linear scale with bounding tics.

Close inspection correctly suggests that the number of quantile tics and regular spacing for the quantile labels drives the piecewise space allocation. Figures 2a and 2b suppress some of regularly spaced labels. The additional labels are useful to those interpolating values but begin to make the plot appear complicated. Cognitive studies may suggest an appropriate balance.

The Figure 2a represents a theoretical truncated normal distribution. Figure 2b represents *pq* values computed from probability sampling in an EMAP (EPA Environmental Monitoring and Assessment Program) study. Since the current emphasis is the graphical design, Figure 2b omits the context and measurement units. However, note that the reference lines show considerable angular variation. Large changes in the

quantiles sometimes correspond to small changes in probability. The largest quantile class shows considerable skewness. More resolution can be provided by selecting the 97th percentile as the bounding quantile. Of course this would hide the information about the higher percentiles.

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 *pq* piecewise linear plot includes more detail without taking up much space. The plot provides a nice compromise between simple legends and extensive *pq* tables.

Closing Remarks

With an already burdensome variety of methodological alternatives, new methods proposed for use should be demonstrably superior to existing alternatives in some significant domain. Newness is not sufficient. (Our interpretation of Pregibon's Razor). In this article we suggest that the *pq* density plot and the piecewise linear plot have sufficient merit to warrant serious consideration.

The ability of plotting methods to generalize is also an important consideration. Can one compare two distributions? What would a parallel *qq* plot look like? How does one add confidence bounds? First thoughts might be that the comparison of Cartesian coordinate curves is so effective that there could not be a viable competitor, but is it so? Cleveland (1985) notes that humans do not judge distances between curves in the correct vertical direction but rather assess distances in a direction roughly normal to the curves. Common confidence lines for CDF plots can be very deceptive. Just maybe there is a better representation for some tasks but that is a topic for another article.

Readers can obtain Splus functions and example script files to conduct their own evaluations or adapt methods to their own applications. Use anonymous ftp to `galaxy.gmu.edu` and look in directory `/pub/submissions/pq`.

Acknowledgements

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

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

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

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.

Carr, D. B. and A. R. Olsen. 1995. "Representing Cumulative Distributions With Parallel Coordinate Plots." Technical Report No. 115. Center for Computational Statistics, George Mason University, Fairfax VA.

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.

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

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

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

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

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

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.

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

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

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

Daniel B. Carr
George Mason University
dcarr@voxel.galaxy.gmu.edu
and
Anthony R. Olsen
U.S. EPA
tolsen@heart.cor.epa.gov