TOPICS IN SCIENTIFIC VISUALIZATION

Plot Production Issues and Details

Background on the Column

In the last issue this column was initiated with an article about the production of stereo plots. We will continue to focus on plot production issues and techniques. Some topics will arise from my *exploratory data analysis* and *scientific visualization* classes. Hopefully the readership will suggest additional topics for development from these two general areas.

Scientific Visualization

Developments in graphics-workstation technology have had a major impact on the field of scientific visualization and should be exploited in the production of statistical graphics. Thus graphics-workstation topics will be fair game for this column. For example, last issue I promised a column on alpha-blending as part of the natural extension to translucent-stereo density plots. While that column is postponed pending scheduling of newsletter color production issues, the topic exemplifies my interest in utilizing advanced technology to facilitate understanding through visual representation. One goal of this column is to encourage cross fertilization between the areas of statistical graphics and scientific visualization.

Statistical graphics involves statistical modeling and the visual representation of central structure, residuals and uncertainty. At least in one interpretation statistical graphics are visual aids for the human endeavor of statistical visualization. Concern about uncertainty as understood through distributional summaries helps distinguish statistical visualization from other areas of scientific visualization.

> ***Statistical graphics involves statistical modeling and the visual representation of central structure, residuals and uncertainty.***

Statistical models define what is meant by central structure and residuals and provide a basis for obtaining distributional summaries. Statistical modeling is an art form that is beyond the scope of this column. This column focuses on plot production details that facilitate the representation of statistical modeling results. Statistical modeling alternatives and issues are raised only in passing. However the importance of modeling software often ties plot production to statistical packages despite the attractiveness of other scientific visualization software.

I have chosen to provide implementations in Splus since it provides a wealth of modeling tools and a programming language sufficient for the production of most static graphics. The Splus code for the examples in the column will be available by anonymous `ftp` from `galaxy.gmu.edu` in the directory `/submissions/eda`. Making the code available allows the procedures to be described in outline form. Hopefully the outline and comments will be of interest to devoted users of other packages. In some cases it may be relatively easy to adapt the provided code to other environments. Please note that the code has been developed in problem solving mode rather than as a polished product. Gentle notes about improvements will be appreciated.

While plot production details are the primary emphasis, the column will occasionally touch on educational materials useful in exploratory data analysis and scientific visualization classes. For example I suspect my list of favorite videos will be of interest. I would like to learn about materials that others are using so I will encourage others to collaborate on education materials columns.

Smoothed Cancer Rates and Hexagon Mosaic Maps

Recently Linda Pickle of the National Center for Health Statistics (NCHS) asked me to contribute a hexagon mosaic map of smoothed colon-cancer mortality rates for use in a NCHS project in evaluating the merits of different map styles. I have asked Linda to provide background about the NCHS project for this column. I then outline the steps followed in producing the hexagon mosaic map and raise some issues that warrant further attention.

Background On Smoothed Mortality Maps

Maps of cancer mortality rates published by the National Cancer Institute in the past (Mason 1975, 1976; Pickle 1987, 1990) turned out to be very successful “visualization tools” for public health researchers. These maps identified cancer “hot spots” and geographic time trends in the U.S. data that had not been noticed before in tabular publications of the mortality statistics. Follow up studies designed to determine the reasons for high rates in particular regions led to such important discoveries as the link between mouth cancer and snuff dipping and lung cancer and exposure to asbestos through shipyard work during World War II.

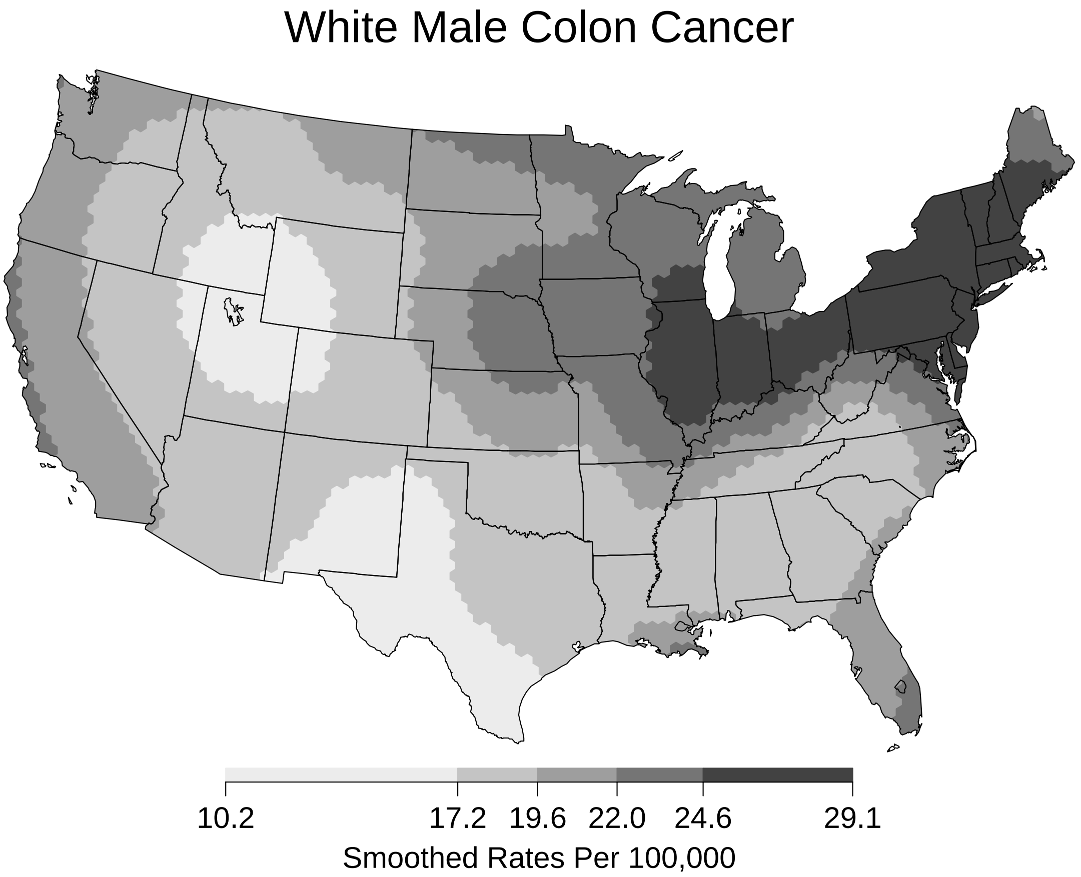

Figure 1

Hexagon Mosaic Map of Smoothed Mortality Rates. The map was produced using Splus, a product of Statistical Sciences, Inc. While Splus comes with a U.S. boundary and Becker and Wilks (1992) have provided map projection and other functions, the particular boundary files used here were obtained from Atlas Pro and are used with the permission of Strategic Mapping, Inc.

Maps of cancer mortality rates published by the National Cancer Institute in the past turned out to be very successful "visualization tools" for public health researchers.

The second atlas series showed the mortality data by time as well as place. Although regional differences in mortality rates seemed to be diminishing over time for many types of cancer, new "hot spots" appeared during the 1970s for several of the major cancers. Because of the success of these atlases, NCHS is planning a mortality atlas of leading causes of death, not limited to cancer. NCHS is the federal agency responsible for collecting and publishing information from all U.S. death certificates.

The problem for earlier atlas designers was how to produce a reasonable looking map overall. For example, it took the combined effort of two federal agencies to produce the hardcopy output for Mason's atlases. Now that mapping software is widely available and a standard desktop PC is powerful enough to process the large mortality data files, the problem we face today is how to choose from the many mapping options available. NCHS has funded several cognitive studies to date to test how geographic patterns are perceived when the underlying data are presented using various map styles. Because NCHS must map the entire U.S. at a small area level (at least 500 geographic units), some maps styles, for example those that use large area symbols such as framed-rectangles, are not feasible.

The goal is to design a map that will present the geographic patterns in the underlying data with the least amount of distortion or perceptual bias. The map should also be able to answer several types of questions typically asked by the viewer: (1) what are the general geographic patterns of rates (e.g. where are the rates high?); and (2) approximately how high are the rates in a certain area? It may be necessary to use a different map style to answer each of these questions.

The goal is to design a map that will present the geographic patterns in the underlying data with the least amount of distortion or perceptual bias.

NCHS has been experimenting with different methods of smoothing the mortality data. That is, if at least some of the random variation in the data can be removed, the broad geographic patterns should be more apparent in the map. Assigning the mortality rate to a single point within an area allows the data to be smoothed and represented by symbols at lattice points thus ignoring the original (usually irrelevant) political boundaries. Because the task is to map events among people a more appealing approach is to assign the rate to the population centroid rather than the geographic centroid, but populations centroids were not available when NCHS created test files for distribution.

For the latest experiment, NCHS provided directly age-adjusted mortality rates for colorectal cancer among white men during 1980-89. The basic geographic units were Health Service Areas (HSAs). The 802 HSAs are groups of counties defined according to where residents obtained their hospital care (Makuc 1991). NCHS is preparing a poster for this year's ASA annual meeting that will summarize the reaction of study participants to a number of map styles, including several smoothed maps.

Production of a Hexagon Mosaic Map

Figure 1 shows a hexagon mosaic map of smoothed colon-cancer mortality rates. The basic hexagon mosaic map is a choropleth map composed of hexagons. Carr, Olsen, and White (1992) used the hexagon mosaic map to represent sulfate deposition and trends and discuss possible merits of this type of map relative to closely related square mosaic maps and pseudo- color contour maps. Ultimately the representational form needs to be evaluated in perceptual studies and the NCHS study provides an all too rare opportunity for evaluation of map styles.

The production and interpretation of maps cannot be divorced from the application. Two differences between the sulfate map in the original application and the current mortality rate maps warrant mention. In terms of production, the sulfate observations were essentially point data. The supplied mortality rates were area-based estimates. Assigning an area-based rate to a point fails to capture finer scale variation within the area when such exits. The detail obtainable is limited to county level statistics for cause of death information. Aggregation to HSAs attempts to reduces variance by pooling information from "similar" counties. Step three below was not necessary in producing a smoothed sulfate deposition map because the transition from area-based estimates to point estimates was not needed.

In terms of interpretation, a smooth is better accepted when there is reason to believe that the underlying surface is smooth and something other than a flat plane. The sulfate deposition process, when viewed over a long period of time, makes a smooth deposition surface over the U.S. seem quite plausible. In fact a strong west to east gradient can be anticipated based on recorded SO^2 emissions and wind currents. The connection between spatial location and mortality rates is less obvious to those outside the epidemiology community. However spatial location often serves as a surrogate for variables that are "causally" related to mortality rates. In fact most major cancer types exhibit spatial clustering. While some spatial clustering might be happenstance, some patterns such as those in Figure 1 have remained relatively stable over three decades. The spatial deposition surface tends to be interpreted as a summary result while the mortality-rate surface is used primarily to generate hypotheses for further investigation.

The following annotated steps indicate decision points and tasks involved in the production of a hexagon mosaic map of smoothed mortality rates.

Step 1 is to pick a map projection. In the original application, sulfate deposition is conveniently described as an amount per unit area. This strongly favors use of an area preserving map projection. In the cancer mortality rate context, the choice of map projection is not so clear. I chose to stay with an Albers equal area conic projection and this choice is consistent with NCHS practice.

Step 2 is to select a grid resolution and generate a set of hexagon centers so the hexagons will cover the map of the continental United States. Selecting a grid resolution has received little discussion in the literature. Since the hexagons in a hexagon mosaic map have direct visual impact, the choice of resolution is likely more important than in applications in which the grid is hidden by subsequent contouring.

Defining a rectangle that encloses the U.S. and generating a hexagon grid that covers the rectangle is straight forward. The problem is eliminating hexagon center points (centroids) for hexagons that fall completely outside the U.S. boundary. Since the U.S. boundary is a polygon each candidate centroid is tested using a point in polygon algorithm (see Littlefield 1984). Centroids inside the U.S. boundary are accepted. Each centroid outside the U.S. boundary is further tested and accepted if any of its surrounding hexagon edges intersects any of the U.S. polygon edges (see Sproelder and Ulling 1990 for an intersecting segments algorithm) .

Step 3 is to obtain point data for use in traditional model software. The supplied values were HSA centroids in Albers coordinates. Algorithms often available in GIS packages can produce centroids from the polygon boundaries of each area. As indicated above, obtaining centroids based on populations appears preferable for this type of data.

Step 4 is to model the point data. Common modeling approaches for point data include kriging, splines, and polynomial regression. The choice here was to model the mortality rates using local regression (loess). Cleveland, Grosse and Shyu (1990) discuss the modeling options available. The particular options used for Figure 1 include local quadratic modeling, the Euclidean distance option for the independent variables, inverse variance weights and direct modeling of data. The independent variables were the Albers coordinates representing the HSA centroids. Using the actual surface of the earth interpoint distances may be technically more desirable but the potential for improvement seems small compared to the complications involved. The inverse variance rate approximation was based on a Poisson death rate model and a rough estimate of the white male population size. Having population data available facilitates more sophisticated modeling. Cleveland and Devlin (1988) discuss the modeling procedure to assist in picking a smoothing fraction for the local modeling. Figure 1 represents the second plot produced and oversmoothes the rates.

Step 5 is to obtain estimates at the hexagon grid locations obtained in Step 2. In Splus the predict function combines the modeling results with new values of the independent variables to obtain estimates. (An alternative is to obtain the average surface value for each hexagon.) Those experienced in the use of 2-D smoothers are typically concerned about the behavior of the smoother near the boundaries of the modeled data (and near sharp discontinuities). In this application some of the hexagon centroids near the edges of the map lie outside the convex hull of the data so even “extrapolation” is involved. The irregularity of the map border also results in smoothing between regions separated by water. The appropriate measure of distance in such cases depends on the application. While values near the fringes of the map should be interpreted cautiously, the extrapolation problem is a mild one numerically since no point on the map boundary lies very far from some HSA centroid and mild philosophically since the modeled data includes people that live near the boundaries.

Splus supports the notion of a rectangular grid. A more general notion would provide for other lattices bounded by polygons. Algorithms could be developed to correspond to these lattices.

Step 6 is to determine the class intervals. My preference would be to determine the class intervals by the percent of people affected. For example a class boundary might be determined so the hexagons in the highest rate class includes five percent of the population. The boundaries in Figure 1 were supplied to be consistent with the HSA quintile boundaries in other plots.

Step 7 is to pick the color for each class and plot filled hexagons using the colors. Plotting hexagons is straight forward using the polygon plotting function.

Step 8 is optional and represents low population regions. Some analysts find it distressing to see values represented for deserts, lakes and rugged terrain where few people live. One possibility would be to overplot the hexagon cells that have populations below a certain threshold to indicate low populations regions. For example a reduced-size background-colored hexagon might be used. This step requires hexagon binning of detailed population data and was not used in Figure 1.

Step 9 is to clip the hexagons back to the U.S. boundary. Clipping involves overplotting undesired regions with polygons in background color. A clipping trick that seems to work (but may not be formally supported) in postscript is to construct a more complex polygon appending a closed surrounding rectangle to a closed U.S. boundary polygon. At least on one system, plotting this complex polygon in background color overplotted the region between the U.S. boundary and the bounding rectangle. A safer approach is to construct two simple clipping polygons by splitting up the U.S. boundary polygon into two pieces and adding a few points.

Note that postscript processing software differs especially in regard to the number of polygon vertices that can be processed. In Splus version 3.1, I had to invoke an option to handle over 2300 vertices in the U.S. boundary. (Another solution is to use a generalized boundary with fewer vertices.) The postscript previewer on my workstation would not handle that many vertices and stopped. Two different postscript printers had no problems with the postscript file. (I have also observed grey-level texture patterns to vary from device to device.)

Step 10 is to add detail such as state boundaries, islands and lakes. Given the boundaries, this simply involves plotting polygons.

Step 11 adds a legend. The current legend involves plotting polygons, lines and text so was straight forward to produce albeit tedious the first time. Legends can provide additional information but that is a candidate topic for another column.

In summary, the production of a hexagon mosaic map is straight-forward, given boundary files and estimates on a hexagon lattice. Software that provides for filling polygons and plotting text should allow the production of such maps. Care needs to be given to the tasks of modeling the data and obtaining good estimates near the map boundaries. Several map enhancements are possible including marking regions with low populations and adding information to the legend. The hexagon mosaic maps fared well in the NCHS initial evaluation. At the very least the hexagon mosaic map provides one more alternative in the arsenal of tools for representing smoothed data on maps.

Research related to this article was supported by NSF under grant no. DMS-9107188 and EPA by EPA under cooperative agreement no. CR820820-01-0. 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

Becker, R. A. and A. R. Wilks. (1993), "Maps in S." AT&T Bell Laboratories Statistics Technical Report 93-2. Murray Hill, New Jersey

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*, **10**(4), 228–236, 271.

Cleveland, W. S. and Devlin, S. J. (1988), "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *Journal of the American Statistical Association*, **83**, 596–610.

Cleveland, W. S., E. Grosse, and W. Shyu. (1990), "Local Regression Models," in *Statistical Models in S*, Eds. J. Chambers and T. Hastie. Wadsworth & Brook/Cole. Pacific Grove CA. pp 309–376.

Littlefield R. J. (1984), "Basic Geometric Algorithms for Graphic Input", In *Computer Graphics '84: Proceedings of the 5th Annual Conference and Exposition of the National Computer Graphic Association, Inc. (Vol 2)*, Fairfax, VA: National Computer Graphics Association, pp. 767-776.

Makuc, D. M., B. Haglund, D. D. Ingram, J. C. Kleinman and J. J. Feldman. (1991), "Health Service Areas for the United States", *Vital Health Statistics*, **2**, 112.

Mason, T. J., F. W. McKay, R. Hoover, W. J. Blot and J. F. Fraumeni, Jr. (1975), *Atlas of Cancer Mortality for U.S. Counties: 1950–1969*. Washington, D.C.: USGPO, DHEW Publ. No.(NIH) 75-780.

Mason, T. J., F. W. McKay, R. Hoover, W. J. Blot and J. F. Fraumeni, Jr. (1976), *Atlas of Cancer Mortality Among U.S. Nonwhites 1950–1969*. Washington, D.C.: USGPO, DHEW Publ. No.(NIH) 76-1204.

Pickle, L. W., T. J. Mason,, N. Howard, R. Hoover, and J. F. Fraumeni, Jr, (1987), *Atlas of U.S. Cancer Mortality Among Whites: 1950–1980*. Washington, D.C.: USGPO, DHHS Publ. No. (NIH) 87-2900.

Pickle, L. W., T. J. Mason,, N. Howard, R. Hoover, and J. F. Fraumeni, Jr, (1990), *Atlas of U.S. Cancer Mortality Among Nonwhites: 1950–1980*. Washington, D.C.: USGPO, DHHS Publ. No. (NIH) 90-1582.

Sproelder, H.J. W. and F. H. Ulling. (1990), "Two-Dimensional Clipping: A Vector-Based Approach", in *Graphics Gems*, Ed. A. Glassner. Academic Press, Inc. New York, pp 121–128.

Daniel B. Carr
George Mason University
dcarr@galaxy.gmu.edu

Linda W. Pickle
National Center for Health Statistics
lwp0@nch09a.em.cdc.gov

NET SNOOPING

Alex and Anonymous FTP

Most internet users are aware of at least one or two sites that offer an anonymous FTP service—that is a repository of information that any user can access via FTP. The service usually works by allowing the user to login as "anonymous" and generally sending their e-mail address as a password. Very few users are aware of all of the different FTP servers available to them. There are several ways of navigating through the wealth of available information.

One common tool for searching through anonymous FTPable material is the `archie` program. Archie is available from many good FTP archives, including my favorite one, `gatekeeper.dec.com`, in the `pub/net/infosys/archie` directory. Once one