than 90 or 95 percent white. Of course, estimates for the somewhat smaller, more narrowly defined white, non-Hispanic segment of the population would have yielded more meaningful minority clusters, based on ethnicity as well as race.

> ... ***marketing firms and the producers of "Beverly Hills 90210" conveniently equate postal codes with income and status,*** ...

But better demographic measures alone would not have made the results meaningful because Zip Code areas are neither census tracts nor neighborhoods. Although marketing firms and the producers of "Beverly Hills 90210" conveniently equate postal codes with income and status, Zip Code boundaries reflect the local geographic organization and operation of the U.S. Postal Service, not the boundaries of homogeneous socioeconomic communities. And even if the Postal Service had deliberately sought racially distinct postal zones, Zip Code areas generally are much larger than census tracts and inherently more diverse. (Although an accident of postal geography might let some Zip Code areas reflect comparatively large minority neighborhoods in major cities, postal zones are more likely to be racially mixed than segregated. Additional information is needed to tell whether a racially mixed Zip Code reflects, for example, a uniform zone with segregated housing or the juxtaposition of a white ethnic enclave and an impoverished black neighborhood.) Moreover, the comparatively large size of Zip Code areas allows a substantial separation between residential and industrial neighborhoods, and between homes and toxic dumps. While residents of rural areas and small cities relying on well water might be highly apprehensive about groundwater contamination anywhere within their Zip Code areas, the water supply systems of large metropolitan areas with substantial minority populations typically rely on aqueducts and reservoirs, not on local aquifers vulnerable to a toxic landfill a block or even a mile away.

Zip Code Convenience

Why then is Zip Code information used so widely in marketing studies and advertising campaigns? Because of the obvious and straightforward link between demographic data and potential buyers or voters. If a study reveals, for example, that well-heeled Republicans account for 70 percent of the households in a Zip Code area, campaign literature mailed to all addresses in the zone will most certainly reach a high proportion of potential voters likely to support a candidate advocating traditional family values and lower top-bracket tax rates. But in the NLJ's analysis, this link was missing. Simply put, aggregated demographic data reported by five-digit ZIP Codes reveal little about the people (or the land use, for that matter) in the immediate vicinity of point locations.

> ... ***the comparatively large size of Zip Code areas allows a substantial separation between residential and industrial neighborhoods, and between homes and toxic dumps.***

Although environmental racism is real and reprehensible, five-digit Zip Codes are a poor basis for a broad, systematic investigation of racist practices in either environmental enforcement or the cleanup of Superfund sites. A credibly thorough geographic analysis requires more detailed information based on smaller spatial units and identified links between toxic dumps and drinking water.

Mark Monmonier
Syracuse University
`mon2ier@mailbox.syr.edu`

TOPICS IN SCIENTIFIC VISUALIZATION

Constructing Legends For Classed Choropleth Maps

by Dan Carr

Constructing map legends is a statistical graphics topic worthy of attention. Map legends can provide a distributional summary for a variable represented on a map. A distributional summary augments the map's spatial information and provides the reader with a few key numbers to remember. As an example, researchers and politicians may be concerned if one of "their" counties had a high death rate due to a specific type of cancer that is environmentally or life-style related. The high rate becomes more deserving of study (or more useful for funding leverage) if the rate for the people in this county is above the 95 percentile for the nation. While death rates have a direct interpretation, a population-based comparison provides a useful standard of reference. Thus distributional summaries should be considered as part of the map construction process.

A great deal can be learned about maps legends by examining the options available in GIS packages and by looking at publications. For example Goldman

(1991) provides numerous county-based choropleth maps showing death rates and environmental hazards. The maps show counties with high death rates or high potentials for exposure and include both density and percentile legends that are based on the number of counties involved. The use of distributional summaries based on the number of political regions is common and often convenient. The distributional summaries promoted here answer questions about percentage of people or percentage area. These differ for summaries that count the number of political regions. Figure 1 provides an example.

The 802 regions represented in the map are health service areas. Health service areas (HSAs) are either counties or aggregates of counties as discussed in the last newsletter article (Carr and Pickle 1993). The spatial patterns of mortality rates are clearly of interest with the higher rates in the Northeast. The legend provides a table lookup capability for the classes of mortality rates. The legend also provides a distributional summary for the percent of the white male population. For example the legend indicates that 50% of white males live in HSAs with rates at or below 22.4 deaths per 100,000.

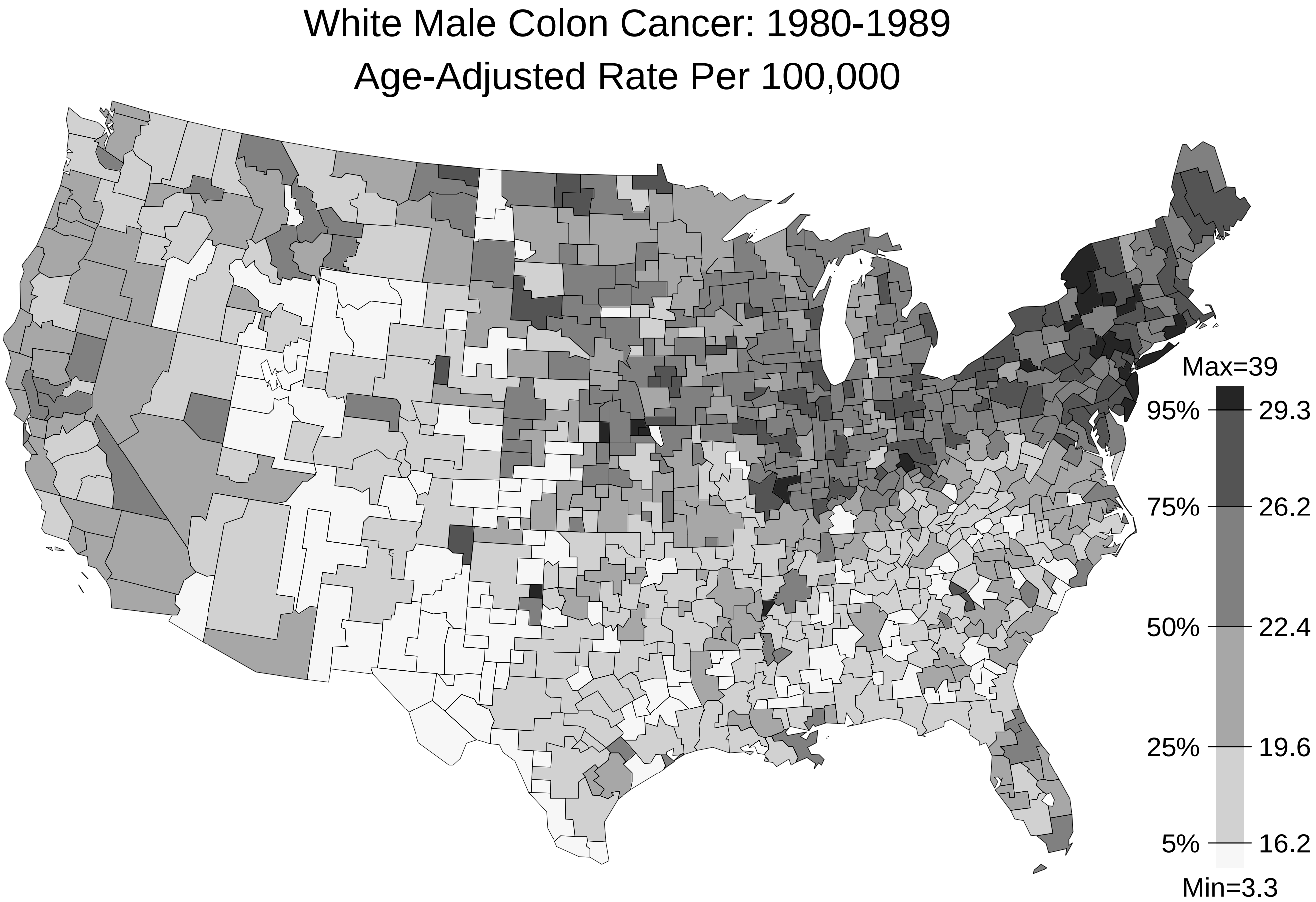
White Male Colon Cancer: 1980-1989
Age-Adjusted Rate Per 100,000
Max=39
95%
29.3
75%
26.2
50%
22.4
25%
19.6
5%
16.2
Min=3.3

Figure 1: Example of a county-based choropleth map.

Figure 2 shows three candidate distributional summaries. All three plots involve sorting the HSA values in increasing mortality rate order before calculating the cumulative values. The left plot is an approximate cumulative distribution for the described population. The center plot gives cumulative percentages for the number of HSAs and the right plot provides the cumulative percentages of map area. The second summary is mostly of interest because it is often used. The area-based summary is often relevant for maps of environmental variables. The area-based summary is interesting here because it provides a rudimentary characterization of what we see on the map. (While the map projection is area preserving, the characterization would be more exact if our visual response were linear with area and unaffected by color interactions.)

A well-known problem with choropleth maps is that the large regions draw visual attention that is not necessarily proportional to the described population. Figure 3 shows the difference in percentages between the area-based and population-based cumulative estimates in Figure 2. The striking result demonstrates once again the importance of plotting the difference between curves rather than visually estimating the difference. The positive percents suggest that after putting the values in mortality-rate order, the large area HSAs are encountered sooner than the high population HSAs. This suggests a relationship between mortality rate and population density. When the population percents define the class intervals, visual attention drawn to the relative areas with different shading can have a population density based interpretation.

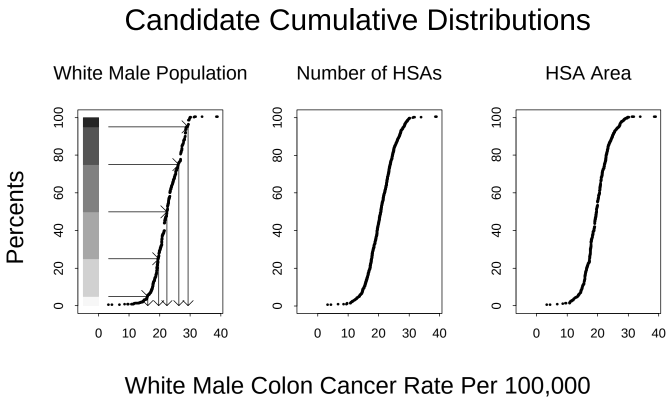

Figure 2: Example cumulative summary distributions.

Consider viewing a series of maps that use the same percents to define class intervals. A few maps may have an unusually small total area devoted to the high rate class and an usually large total area devoted to the low rate class. For gray scale maps this changes total amount of reflected light and acts as a cue. Defining the class intervals symmetrically with respect to population percents allows the meaningful class area comparisons within a map. This comparison can be done for the pair classes in gray scale maps such as Figure 1.

> ***The striking result demonstrates once again the importance of plotting the difference between curves rather than visually estimating the difference.***

For example, the lowest rate class covers a much larger area than the highest rate class. The comparisons are easier when a symmetric color scheme is used like the one described below. Of course directly plotting mortality rate versus population density is better than relying on visual estimates of class area differences and other variables need be considered than surrogates like population density and spatial position. The point is that the area devoted to the classes can be suggestive.

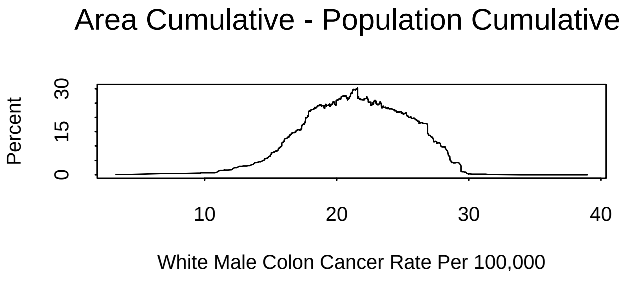

Figure 3: Difference in percentages between area-based and population-based cumulative estimates.

The legend in Figure 1 shows both population percents and mortality rates as class boundaries. The mortality rates are quantiles determined from the cumulative distribution. The arrows in the left plot of Figure 2 illustrates the process of going from selected percents to the corresponding quantiles. A convenient approach is to linearly interpolate between the points on the cumulative distribution rather than treat the distribution as a right continuous step function. Thus the estimated quantiles may not produce the intended percentages when used as class boundaries. However, using more accurate but non-standard percentages complicates the description. Unless there is big discrepancy, simplicity suggests emphasizing that the quantiles are approximate.

Simple Design and Boxplots

When designing map legends it helps to pay attention to graphics methods that have been successful in the past. The boxplot has been one of the few modern statistical plots that has worked its way into elementary statistical texts and into use by applied scientists. The boxplot with its emphasis on approximate .25, .5 and .75 quantiles recognizes some important factors in its design. First it uses a simple standard (.25, .5 and .75) that is easy to accept. Second, the standard relates strongly to important concepts of central tendency, distributional spread, and assessment of symmetry. Third, the summary focuses mental attention on a few values that can be used for making comparisons.

Emphasizing a few values for mental comparison is an important principle. The human short term memory can only handle about 7 ± 2 units of information at one time. (Depth of thought may relate more to what people use as units of information than to the ±2). Ehrenberg (1981) argues persuasively that short term memory considerations should be used in the design of tables. For example an ordinary person can divide a two digit number by a two digit number and have room to store an approximate two digit answer. Most people have difficulty when they try to ratio two three-digit numbers. What happens in the graphical environment is somewhat different because the graphic can be used for rapid mental refresh. However one might conjecture that people will withdraw from map reading if there are more than seven or so obvious and equally important classes or layers of information.

Color Scales

The legend in Figure 1 uses six classes with internal boundaries determined approximately by 5, 25, 50, 75, and 95 percentiles. While humans can easily distinguish many more than six gray levels, Figure 1 appears complicated. Part of difficulty relates to the dot-based representation of gray, part relates to the spatial variability that provides a changing background against which to judge color and part relates to using as many as six "equal" classes. When full color is available the map can be made to appear simpler by using shades of red for high rates and shades of blue for low rates. This "grouping" of information has only two equal classes at the top and three ordered classes nested inside. The red and blue colors can be ordered both in terms of saturation and value. Using low-saturation near-white colors in the middle of the scale eases the transition from blue to red. Putting the saturated and dark colors at the extremes follows the advice of Eduard Imhof (see Tufte 1990) by devoting relatively little area to saturated colors.

Of course other color scales can be considered, but the above scale is a reasonable start for those who are not red color blind. The verbal description above does not do the color selection justice. For those interested, compressed color postscript files are available by anonymous `ftp` to `galaxy.gmu.edu` and stored under `submissions/eda/maps`. The directory also contains the Splus commands files and the data used to produce the maps. Splus users can easily modify the colors and experiment with legend variations.

Variations

Legend variations are worth considering. Those who are intensely studying the phenomena may want to see the full cumulative distribution so they can read the percentage for any mortality rate. The Figure 1 legend shows only selected values both to save space and keep the legend simple. The legend scaling is linear in percentages rather than mortality rates to provide programming convenience and generality. If the legend scaling were based on mortality rates, the quantiles corresponding to the standard percentiles could be so close that they would overplot. For visual communication a linear scale in terms of mortality rates would also be helpful.

When full color is available the map can be made to appear simpler by using shades of red for high rates and shades of blue for low rates.

Many map variations are worth considering. Perhaps the most suggestive and elegant map in the map directory cited above is that showing the extreme residuals from the smooth. These local discrepancies from the smooth can be very useful for hypothesis generation. However new topics like exploration of spatial residuals and disaggregation approaches to smoothing using high resolution population data deserve separate consideration.

Acknowledgements

I thank Linda Pickle of the National Center for Health Statistics for discussions concerning the data and Tony Olsen of the U.S. EPA for stressing the importance of distributional summaries. Research related to this article was supported by NSF under grant No. DMS-9107188 and 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

Carr, D. B. and L. W. Pickle. 1993. "Plot Production Issues and Details: Smooth Cancer Rates and Hexagon Mosaic Maps." *Statistical Computing & Statistical Graphics Newsletter,* Vol. 4, No. 2, pp. 16-20.

Ehrenberg, A. S. C. 1981. “The Problem of Numeracy.” *The American Statistician,* Vol. 35 No. 2, pp. 67-71.

Goldman, Benjamin A. 1991. *The Truth About Where You Live, An Atlas For Action on Toxins and Mortality.* Times Books, New York.

Tufte, Edward R. 1990. *Envisioning Information.* Graphics Press, Cheshire, Connecticut.

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