Interactive Linked Micromap Plots And
Dynamically Conditioned Choropleth Maps

By Daniel B. Carr[1,2], Jim Chen[1], B. Sue Bell[2], Linda Pickle[2], Yuguang Zhang[1]

Affiliations: George Mason University[1], National Cancer Institute[2]
Email: dcarr@gmu.edu, jchen@cs.gmu.edu, bellsu@mail.nih.gov, picklel@mail.nih.gov,
            yuguang_zhang@freddiemac.com
URL: www.galaxy.gmu.edu/~dcarr/ccmaps

Abstract

This paper introduces interactive extensions to two recently developed templates for displaying geospatially-indexed estimates. The first template, linked micromap plots, links small generalized maps with statistical panels that describe regions. Research centered at the National Cancer Institute addressed the task of communicating state and county cancer statistics and tailored this template to show estimates, confidence intervals, and Healthy People 2010 target values. The research also integrated interactive options, such as variable selection, sorting, fixed header scrolling, mouse tips, enlarged dynamic map views and drill down, in a Java applet. This template has fared well in early usability tests. The second template, called conditioned choropleth maps, seeks to improve hypothesis generation about the spatial patterns shown in a classed choropleth map. Since variation of a study variable is often related to known risk factors, the template provides a way to control for the known variation. This paper describes dynamic sliders that change class boundaries for a study variable and that partition regions into a 3 x 3 layout of maps based on values of two risk factors. Highlighted regions in each map are more homogeneous with respect to both risk factors. Comparisons across maps and spatial patterns within maps provide the basis for generating hypotheses. The JAVA application shareware also includes dynamic statistical annotation and QQplots for comparing distributions

1. Introduction

This paper describes two recently developed templates for displaying geospatially-indexed estimates: linked micromap (LM) plots and conditioned choropleth (CC) maps. Previous research for federal agencies in displaying statistical summaries led to development of these templates. The primary purpose of this paper is to present recent interactive and dynamic extensions of the two templates that are products of recent digital government research centered at the National Cancer Institute (NCI). The motivating application is web distribution of state cancer information. However, the mapping methodology has broad applicability and transfer to other federal agencies associated with this digital government research project has started.

Two common goals in developing these mapping templates were to integrate more statistical information into a display than a traditional choropleth map and to provide for more rapid assessment of statistical and spatial patterns than would be provided by a table. The particular layout and integration of information makes these templates distinct from previous graphical templates. The addition of interactive and dynamic methods further distinguishes these templates from historical approaches.

While the templates share some goals, they differ in their usage and implementation. The primary purpose of LM plots is to communicate patterns in statistical summaries. Its interactive implementation as a Java applet provides different views of previously prepared statistical

summaries. The primary purpose of CCmaps is hypothesis generation. This Java shareware is implemented as an application to reside on the researcher's own computer and thus have ready access to the researcher's own data. Of course there is no restriction on how investigators use the templates. Carr et al. [3], for example, describe LM plots in a data mining context. Different usage can motivate different implementations.

The paper describes the two templates in turn. Section 2 concerns LM plots and Section 3 describes CC maps. Since these templates are little known, this paper briefly describes the basic elements of the templates as illustrated in Figures 1 and 2. This research represents several years of effort to transform the way federal agencies display their statistical summaries. Section 4 closes by suggesting why these two templates are likely to be used by many agencies.

2. Linked Micromap (LM) Plots

The primary purpose of LM plots is the communication of geospatially-indexed statistical summaries. Figure 1 illustrates the key features of the LM plot template. The left column contains micromaps, the second column contains study unit names, while the third and fourth columns contain statistical panels. As indicated in [1, 2], a LM plot has at least three types of panels (micromaps, names and statistical panels). These panels can take various forms. For example, a "micromap" can be any spatial representation from a human body caricature to a communication network.

Published LM plots [1,2,3,4,5,6] have shown many types of statistical panels. They include dot plots with confidence bounds, time series plots, before and after plots, box plots, scatter plots, smoothed rank order scatterplots, bivariate box plots, and even tiny bars in panel that show 159 variables per region. Some of the box plots have summarized values computed from multiple satellite images each involving 8 million pixels. The only constraints on the statistical panels are creativity and available plotting space.

Figure 1 provides the basis for indicating some differences between LM plots and classed choropleth (colored region) maps. The statistical panels are dot plots with confidence bounds. In contrast to the choropleth maps, the dot plots show the statistical estimates with an encoding that has the highest perceptual accuracy of extraction, i.e., position along a scale [7]. In contrast, classed choropleth maps degrade continuous estimates by transforming them into an ordered categorical variable with a few classes. Then these classes are represented with color, a poor encoding for an ordered variable. While Figure 1 shows confidence bounds as an indication of estimate reliability, reliability is not usually shown in a choropleth map. Choropleth maps have merit in terms of providing a quick overview, but there are many merits to using LM plots.

Figure 1 is a variation that illustrates the inclusion of two reference values. Dashed lines at the edge of the green regions indicate the U.S. Healthy People 2010 targets. The black dashed line is the U.S. reference value. When the confidence bounds do not reach into the green region there is low probability that the states have reached the target.

Additional key features of LM plots are sorting, perceptual grouping, and linking of multivariate descriptors. The study units in Figure 1 are states. These states are sorted by lung and bronchus cancer mortality rates that appear in the third column. After sorting, the design partitions states into small perceptual groups. States are grouped into fives (with one exception). Higher-level grouping creates three blocks of states distinguished by mortality rates above, equal to, and below the median. Distinct hues distinguish the five states in each group. The same hue links a state name, its representation in the micromap, and its estimates in statistical panels. Vertical position also links the name and estimates in most LM plots.

Recent research in developing an interactive Java implementation has centered at NCI where the State Cancer Profiles project seeks to use data visualization on the Internet to disseminate statistics useful in cancer control planning. The States Cancer Profiles project uses LM plots to show two variables: a mortality rate such as lung cancer mortality and a related risk factor, such as cigarette smoking rates. NCI research is evaluating many interactive options for altering the display. The more obvious options include selecting different variables for display and sorting of regions (triangle icons appear above the columns.) Additional features include mouseovers and linked blinking, color selection, a fixed header scroll bar, enlargement for micromaps, and drill down to see the county statistics for any selected state. The current work at NCI on LM plots is in progress and not yet approved for public release. A Java applet showing trial interactive extensions to LM plots using sample statistical data is available at http://graphics.gmu.edu/~xwang/cancer4/index.html.

The goal at NCI is to develop, evaluate, improve and deploy methodology for communicating cancer statistics to state epidemiologists and other public health professionals. LM plot usability assessment is ongoing. The first rounds were based on expert opinion and preliminary testing of volunteers at meeting of the American Cancer Society's regional planners. The feedback has raised some issues. For example, when using a low resolution monitor, the full display extended past the screen and a second slider appeared. Some of those engaged in the assessment task of finding values for specific counties did not realize that some panels were off screen and they did not find the values. There is a learning curve for using LM plots and some addressable difficulties have been uncovered. Nonetheless the small sample of volunteer subjects ranked LM plots above the other graphics proposed to disseminate the State Cancer Profiles. More extensive evaluation is planned after all the graphics are improved based on the feedback.

3. Conditioned Choropleth (CC) Maps

The purpose of CCmaps is to help researchers generate sharper hypotheses about observed spatial patterns. So far there is only one paper [4] that describes CCmaps in much detail. An example has appeared in an encyclopedia article [6] and more publications should follow.

National Center for Health Statistics (NCHS) and Environmental Protection Agency (EPA) applications motivated the development CCmaps. The particular context considered here comes from NCHS, the study of mortality rates. Before conjecturing about the spatial patterns of mortality rates, it is important to produce maps that control for suspected risk factors. In epidemiology, common practice makes separate plots by race and sex to control for differences in study unit populations. Also, an age-adjusted map or a set of age-specific maps controls for study unit differences in age distribution. However, it is uncommon to see maps where efforts have been taken to control for other risk factors. While sophisticated regression models provide the best means of controlling the variation due to risk factors, CCmaps provides more rudimentary but widely accessible control by partitioning study units into more homogeneous groups based on two risk factors.

CCmaps is no longer just a static template but now a Java application available as shareware from www.galaxy.gmu.edu/~dcarr/ccmaps. The application features newly developed partitioning sliders. CCmaps gently incorporates concepts from statistics. The design encourages the analyst to create, examine and contrast subsets for the purpose of generating sharper hypotheses about patterns in spatially-indexed statistics. This is quite different than the filter and drill down methodology that is so prevalent on the web.

As Figure 2 suggests, CCmaps supports dynamic partitioning of study units into a 3 x 3 layout of maps using partitioning sliders. The study units in Figure 2 are health service areas (HSAs), counties or aggregates of counties based on where people get their hospital care. HSAs highlighted in a panel have risk variables satisfying row and column constraints. The slider at the bottom partitions HSAs into columns based on precipitation. The slider at the right partitions HSAs into rows based on percent of households below the poverty level. The slider at the top partitions HSAs into three color classes (shown as blue, gray and red) based on lung cancer mortality rate. Hypotheses can be about spatial patterns within panels or differences among panels. Note the red in the top right panel that highlights HSAs with high precipitation and high poverty. One hypothesis might be that Southeastern HSAs have higher cigarette smoking rates. However the strong association with precipitation warrants deeper consideration.

Current features of CCmaps include statistical annotation and plots. Sliders show the percent of **people** in each class instead of the percent of study units. In Figure 2, 36 percent of the people reside in the study units shaded blue. The population weighted mean rate for HSAs highlighted in each panel appears at the top right of the panel. For Figure 2's top left panel, the population weighted mean lung cancer mortality rate is 35 deaths per 10,000 white males ages 65-74. Clicking on a panel enlarges it to full screen size and clicking again puts it back. An innovation is a separate 3 x 3 layout view of dynamic QQplots (not shown) that facilitates a nonparametric comparison of distributions.

The development of CCmaps is in a state of rapid progress as of this writing. A pan and zoom widget to support closer inspection of US county examples is in progress. The sliders work nicely for the maps with over 3000 counties, but performance slows in the QQplots due to the extensive calculations needed. The list of planned enhancements and examples is long.

4. Acceptance Considerations

There are many barriers to acceptance of new methodology by federal agencies. The methodology must serve a perceived need or else education must occur to make people aware of the opportunity provided in order to justify a change. The enthusiasm of those promoting new methods often needs to be backed up by usability tests that demonstrate their merits. Expense for the software, education and compatibility with previously published data are other issues. While agencies expenses for routinely used software are recurrent budget items, the budgets for new software and education can be very limited. It is difficult to develop a community of researchers who will use the software if it is expensive or has an unacceptable learning curve.

The prognosis for both templates becoming heavily used by the federal agencies is excellent. A key factor has been the development of the templates to address needs perceived by insightful federal staff. These staff members were already advocates for change. Research developed LM plots for EPA applications. Applications at that Bureau of Labor Statistics, the National Agricultural Statistical Service, NCHS, and the Bureau of Transportation Statistics lead to more variations. The implementation of interactive methods at NCI has been particularly advantageous because NCI has in-house usability assessment staff with expertise not only usability but also in the special requirements of the federal agencies. The connections are in place for rapid sharing of development across agencies and many of the needs are similar.

The software has been implemented in JAVA and can move between agencies without software budgetary considerations. Thus the software has a chance to take hold. In the longer term the use of commercial software version seems advisable. Long term academic and in house federal software development and maintenance support often proves problematic.

Bibliography

[1] DB Carr and SM Pierson. "Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps," *Statistical Computing & Graphics Newsletter*, 1996; **7**(3): 16-23.

[2] DB Carr, AR Olsen, JP Courbois, SM. Pierson, and DA Carr. "Linked Micromap Plots: Named and Described," *Statistical Computing & Graphics Newsletter*, 1998; **9**(1): 24-32.

[3] DB Carr, AR Olsen SM Pierson, and JP Courbois. *Using linked micromap plots to characterize Omernik ecoregions. Data Mining and Knowledge Discovery* 2000; 4:43-67.

[4] DB Carr, JF Wallin, and DA Carr. "Two New Templates for Epidemiology Applications. Linked Micromap Plots and Conditioned Choropleth Maps," *Statistics in Medicine* 2000; 19:2521-2538.

[5] DB Carr. Designing Linked Micromap Plots for States with many Counties," *Statistics in Medicine* 2001; 20:1331-1339.

[6] DB Carr, D. B. 2002. "Graphical Displays," *Encyclopedia of Environmetrics*, Eds. A. H. El-Shaarawi and W. W. Piegorsch, Vol. 2. John Wiley & Sons, pp. 933-960

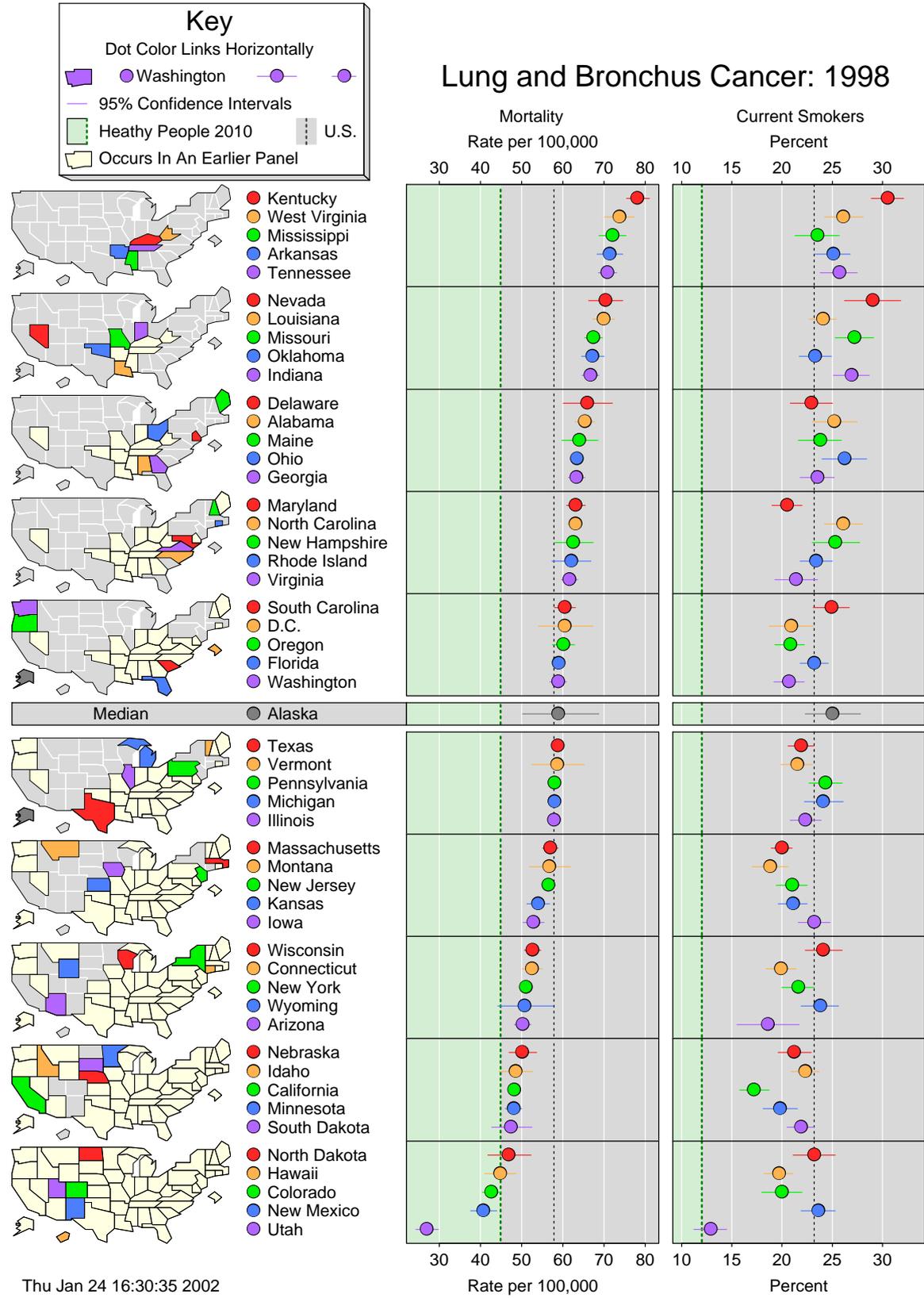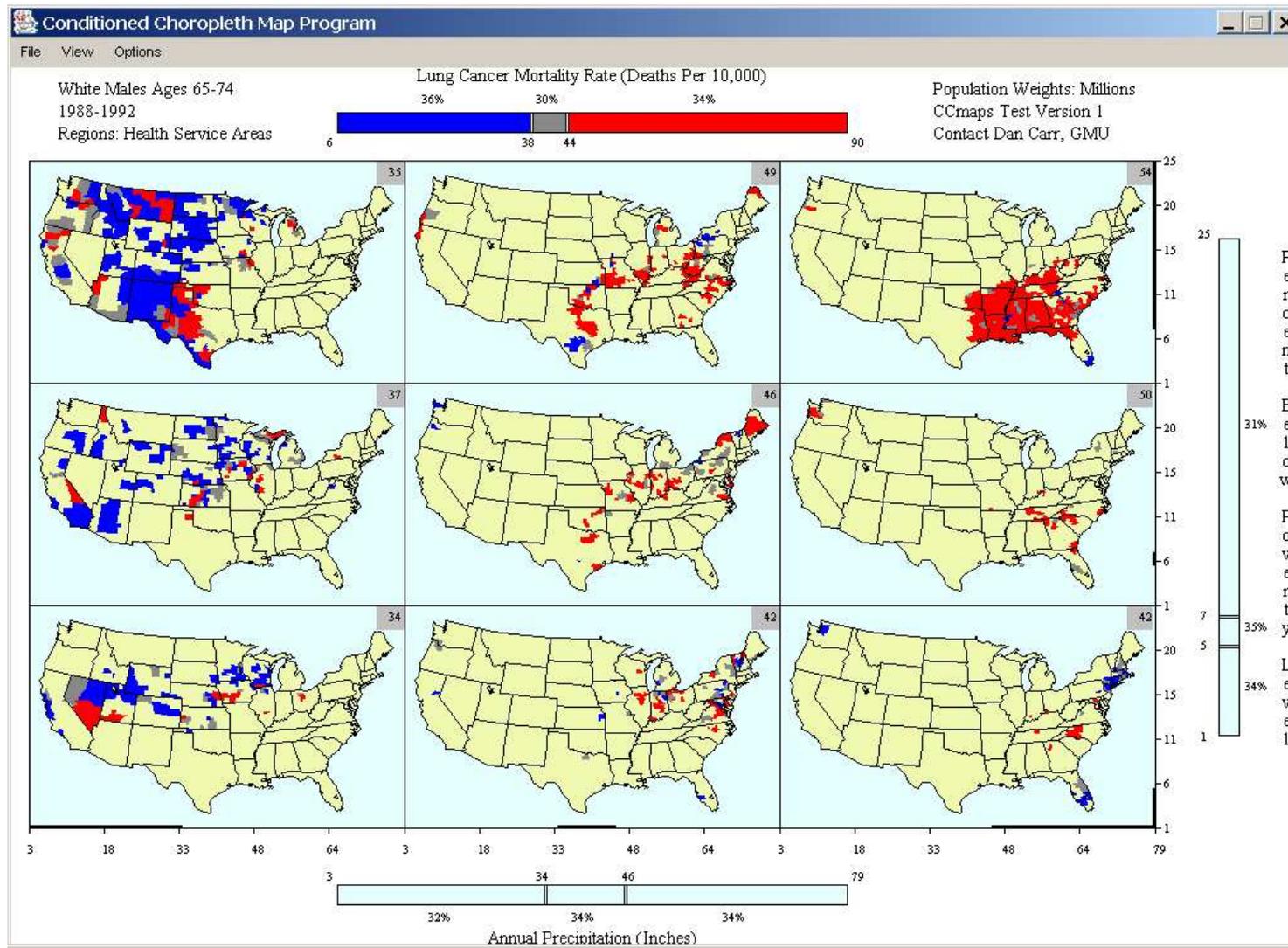[7] Cleveland, W. S. *Visualizing Data*, Summit NJ: Hobart Press. 1993.

Figure 1. A Linked Micromap Plot

Figure 2. A Conditioned Choropleth Map