Journal of Statistical Software

January 2015, Volume 63, Issue 3. *http://www.jstatsoft.org/*

micromapST: Exploring and Communicating Geospatial Patterns in US State Data

Linda Williams Pickle
StatNet Consulting LLC

James B. Pearson, Jr.
StatNet Consulting LLC

Daniel B. Carr
George Mason University

Abstract

The linked micromap graphical design uses color to link each geographic unit's name with its statistical graphic elements and map location across columns in a single row. Perceptual grouping of these rows into smaller chunks of data facilitates local focus and visual queries. Sorting the geographic units (the rows) in different ways can reveal patterns in the statistics, in the maps, and in the association between them. This design supports both exploration and communication in a multivariate geospatial context. This paper describes **micromapST**, an R package that implements the linked micromap graphical design specifically formatted for US state data, a common geographic unit used to display geographic patterns of health and other factors within the US. This package creates a graphic for the 51 geographic units (50 states plus DC) that fits on a single page, with states comprising the rows and state names, graphs and maps the columns. The graphical element for each state/column combination may represent a single statistical value, e.g., by a dot or horizontal bar, with or without an uncertainty measure. The distribution of values within each state, e.g., for counties, may be displayed by a boxplot. Two values per state may be represented by an arrow indicating the change in values, e.g., between two time points, or a scatter plot of the paired data. Categorical counts may be displayed as horizontal stacked bars, with optional standardization to percents or centering of the bars. Layout options include specification of the sort order for the rows, the graph/map linking colors, a vertical reference line and others. Output may be directed to the screen but is best displayed on a printer (or as a print image saved to any file format supported by R). The availability of a pre-defined linked micromap layout specifically for the 51 US states with graphical displays of single values, data distributions, change between two values, scatter plots of paired values, time series data and categorical data, facilitates quick exploration and communication of US state data for most common data types.

Keywords: linked micromaps, spatial patterns, geovisualization, US states.

1. Introduction

Uncovering and displaying geospatial patterns and their associations with demographic, environmental and other factors is difficult due to the multidimensionality of the underlying data. Not only are the basic geospatial data two-dimensional, but the other factors of interest may be interrelated, leading to complexities of their associations with the outcome variable of interest. Linked micromap plots visually link geographic and statistical data and have been shown to be useful for both data exploration and communication of georeferenced data (Carr and Pierson 1996; Carr, Olsen, Courbois, Pierson, and Carr 1998; Carr, Wallin, and Carr 2000; Carr and Pickle 2010). The design is widely applicable and has been used to display, for example, cancer rates and risk factors (National Cancer Institute 2013b), birth defects (Gebreab, Gillies, Munger, and Symanzik 2008), growing degree days and precipitation by ecoregions (Carr *et al.* 1998), and financial data (Blunt 2006).

The linked micromap graphical design uses distinct colors within perceptual groups of geographic units to link each location's name with its statistical graphic elements and map location. Columns of panels show names, statistics and maps. Perceptual grouping of the panels at two different scales of organization facilitates local focus and visual queries. Sorting the geographic units in different ways can reveal patterns in the statistics, in the maps, and in the association between them. This design supports both exploration and communication in a multivariate geospatial context.

The state is often the preferred geographic unit for analysis and display of US data. For data collected routinely across the entire US, state populations are large enough to show stable patterns for many types of outcomes and to alleviate privacy concerns, e.g., potential identification of individuals in health data, yet small enough to show sub-national data patterns. In addition, a map of US states is familiar to most people, even to analysts outside the US For these reasons, data collection agencies and users of their data tend to prefer a state-based display. The boundaries of our small maps have been modified so that the color fill of even the smallest state can be readily identified.

In this paper, we describe an R (R Core Team 2014) package **micromapST** that creates a linked micromap plot specifically designed to easily and quickly display data for the 50 US states plus the District of Columbia (DC) (Carr, Pearson, and Pickle 2013). (For convenience, we will hereafter refer to these 51 units as "states".) The entire plot fits on a single page in portrait orientation to support rapid visual queries. Each column in the design may address a different topic and the information displayed for each topic may consist of one or more statistical variables, a map or a state name label. The graphic elements (glyphs) for each state/topic (row/column) combination may represent a single statistical value, e.g., by a dot or horizontal bar, with or without an uncertainty measure. Two values per state may be represented by an arrow indicating the change in values, e.g., between two time points, or by a scatter plot of the paired data. The distribution of sub-state values, e.g., for counties within each state, may be compared across states by means of box plots. Multiple pairs of state values can be shown as line graphs or scatter plots. The line graphs may represent, for example, a time series, a cumulative distribution or a Q-Q plot of each state distribution relative to the national distribution. Categorical data can be represented as stacked bars. The user can choose the row sort order, column order and the type of glyph used for each panel without having to be concerned about the sizing or spacing of the plot components required to create a publication-quality graphical display. The final layout may be saved in

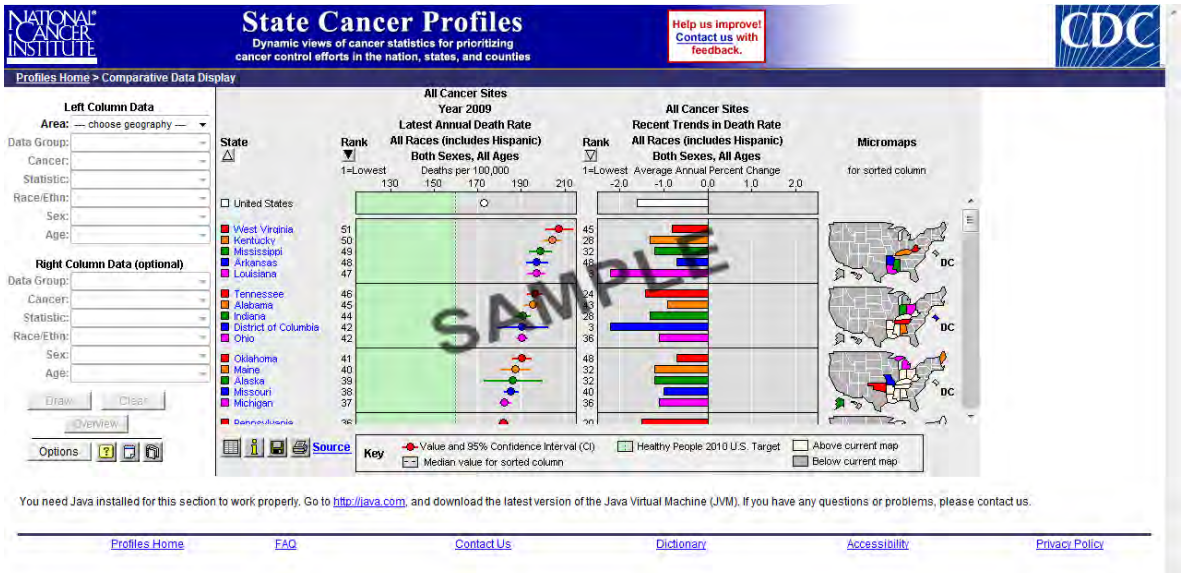

Figure 1: Implementation of linked micromap plots by the National Cancer Institute and the Centers for Disease Control and Prevention.

any file format available in R, including PDF, PNG, BMP, TIFF, JPEG, postscript and SVG, although there are rendering differences among the formats.

Linked micromap plots to date have been produced by customized S-PLUS or custom R code (e.g., in Carr and Pickle 2010). A micromap add-in for SAS JMP is available (Ventura 2012) but only comparative and conditional, not linked, micromap designs are provided. The National Cancer Institute (NCI) developed two Java programs for linked micromap display. The first is an interactive web-based program that displays cancer statistics alongside demographic and cancer risk factor data for states within the US or counties within a specified state (see National Cancer Institute 2013b; Bell, Hoskins, Pickle, and Wartenberg 2006). The aspect ratio of computer displays and the need for header and footer content limits the vertical space available so that, in this example in Figure 1, fewer than half of the states are visible on a single screen. This limitation interferes with the visual comparisons of all 51 states. A printer-friendly layout including all states can be created by clicking the printer icon in the lower left corner of the screen. While this has been a popular method for displaying statistics for 20 common cancers, this Java applet's GUI is specific to the NCI database and cannot be extended to show user-provided data.

The second NCI Java program is more general and can display linked micromap plots for any statistical data with a geographic identifier that can be linked to state, county or other geographies via a `.gen` or `.shp` file (National Cancer Institute 2013a). This program can display glyphs comprised of bars, box plots, dots, dots with confidence intervals and arrows and offers several display options but is designed more for data exploration than to produce publication-quality output. In addition, because this is a standalone Java package, one must use other software to create or modify the data during an iterative exploration of the data.

The R package **micromap** recently released on the Comprehensive R Archive Network (CRAN) by the US Environmental Protection Agency (EPA, see Payton, McManus, Weber, Olsen, and Kincaid 2015) should help to popularize the linked micromap design by embedding the display

capabilities within the widely used and interactive software package R (R Core Team 2014). The current version offers dot and bar glyphs for single values, with and without confidence intervals, and box plots that summarize the distribution of values for each geographic unit. Because this package is based primarily on the R package **ggplot2** (Wickham 2009), it supports construction of other glyphs by the user. The **micromap** package is flexible in that it can utilize any boundary structure and the statistical data frame can be easily constructed in R. However, the price of this flexibility is execution speed and the user must attend to all the details of producing publication quality output, such as refining the margins, spacing, text labels and other graphic elements. Our package, **micromapST**, has pre-specified these details, so that the initial setup for a linked micromap of US state data is quick and easy to do, even without expertise in R coding. By changing the specification of the sort variable or column content, data exploration can proceed quickly and the final linked micromap display will be publication ready.

2. The linked micromap design

The linked micromap design, developed for the EPA, first appeared on a poster in 1995 and was first published in 1996 (Carr and Pierson 1996). A panel layout function used in our new package is a direct descendent of a script that produced linked micromaps for the Bureau of Labor Statistics in 1997 (Carr 1997). The design was first named and described by Carr in 1998 (Carr *et al.* 1998) and further elaborated in 2000 and 2001 (Carr *et al.* 2000; Carr 2001).The early plots included the District of Columbia with the 50 states because that was a requirement for federal clients. These initial designs focused on tasks satisfied by plotting state glyphs along the same row as the state name. Figure 2 illustrates the linked micromap design set up to specifically display the 51 states on a single page for ease of comparison.

Perceptual grouping of the 51 state rows occurs at two scales. There are three large scale groups: a large block at the top of the page with 25 states, a block with one panel containing a single state (typically the median state by some measure) and, symmetrically, a second large block of 25 states at the bottom of the page. Each of the large blocks is further divided into panels of five states each, displayed on a single map, allowing the reader to quickly focus on smaller chunks of the data in the panels.

In this design, color is only used to lead the eye across the row, linking a state name, its glyphs and its map polygon. The color is not an indication of any particular value on the map, as would be the case for a choropleth map. Because of this, the set of five colors can be reused for each panel of five states. The default colors are very distinct hues that can easily be named and distinguished on the small maps. A gray scale palette is available as an option when necessary but it is easier to scan for a named color than for a certain shade of gray.

The use of small multiples of maps and glyphs to facilitate comparisons is consistent with the recommendations of Bertin (1973) and Tufte (1983). By holding the design of each graphic constant (e.g., axis scales, colors), the reader can focus on the data while scanning across the multiple graphics, leading to the most accurate assessment of the statistical values (Cleveland and McGill 1984) and changing patterns across the panels (Tufte 1983). In addition, if all of the small graphics can fit within the field of vision of the reader, i.e., on a single page or display view, then comparisons are more accurate than when the reader must flip between pages or rely on memory of previous images (Tufte 1990). Requiring all graphics to fit on

a single page constrains the vertical space available for representing statistics using position along a scale or length encodings and can lead to overplotting or less than optimal aspect ratios (further discussed in Section 4.4). All graphical designs involve compromise based on prioritizing competing goals. Our highest priority is produce a high-quality image that allows accurate comparisons of US state data on a single page.

Other important design features of the linked micromap plot are the juxtaposition of maps and graphs (Monmonier 1988), and sorting with perceptual grouping to facilitate pattern recognition and image retention in memory (Wainer 1993; Carr and Olsen 1996; Friendly and Kwan 2003). Kosslyn (2006) and others suggest perceptual groups of four or fewer items for comparison. Four states per group would require 13 rows of map panels. The small vertical space for each of 13 micromaps would make it difficult to perceive the color fill for small states. Our design compromises Kosslyn's guidance by using only 10 micromaps, each highlighting five states. The single state panel in the center has no map because of the space limitations. Therefore the design uses a sixth color, black, to highlight this state's location in the maps just above and below the middle panel. The design separates columns of statistical panels with a small amount of white space to reduce the problem of grid line label over plotting from the adjacent panels.

Figure 2 illustrates this design. The leftmost column contains a series of small maps, one for each panel of five states. The US state map is a modification of Monmonier's visibility map (Monmonier 1993) with highly generalized boundaries and enlarged areas of extremely small states. Thus state shapes on the small maps are still recognizable and large enough that their color fills are readily identified. Each map is followed by a list of state names with a rectangle indicating the color of that state on the map for that panel. The next two columns show panels of glyphs for selected statistical information; these glyphs will be discussed further in Section 4. The state rows have been sorted in descending order by the rates in the next panel. The eye is led across the row for each state by the linking color, showing the plotted values and the geographic location of that state. Reading down a column provides a quick visual comparison of the sorted data values in graphic panels or the geographic locations of states with similar ranked values in map panels. Vertical white grid lines lead the eye from a glyph to the common scale labels at the top or bottom of the full graphic. Unlike choropleth maps with each color assigned to a range of values, all states' values are plotted so that differences between states can be seen quickly. The number of columns on the page is limited only by the page width and required resolution for each panel.

In Section 4, we provide cognitive justification for the glyphs available in the package. More details on the research behind the linked micromap design can be found in the book *Visualizing Data Patterns with Micromaps* (Carr and Pickle 2010).

3. Basic setup of data and function call

The package **micromapST** is available from CRAN at `http://CRAN.R-project.org/package=micromapST`. The input data must be contained in a data frame that consists of at least one data column and 51 rows with row names being the state names, two-character postal codes, or two-digit fips codes (`rowNames = "full"`, `"ab"` or `"fips"`, respectively, in the function call). The second input argument to the **micromapST** function is a plot description data frame. This describes the content of each column in the plot. Column specification includes

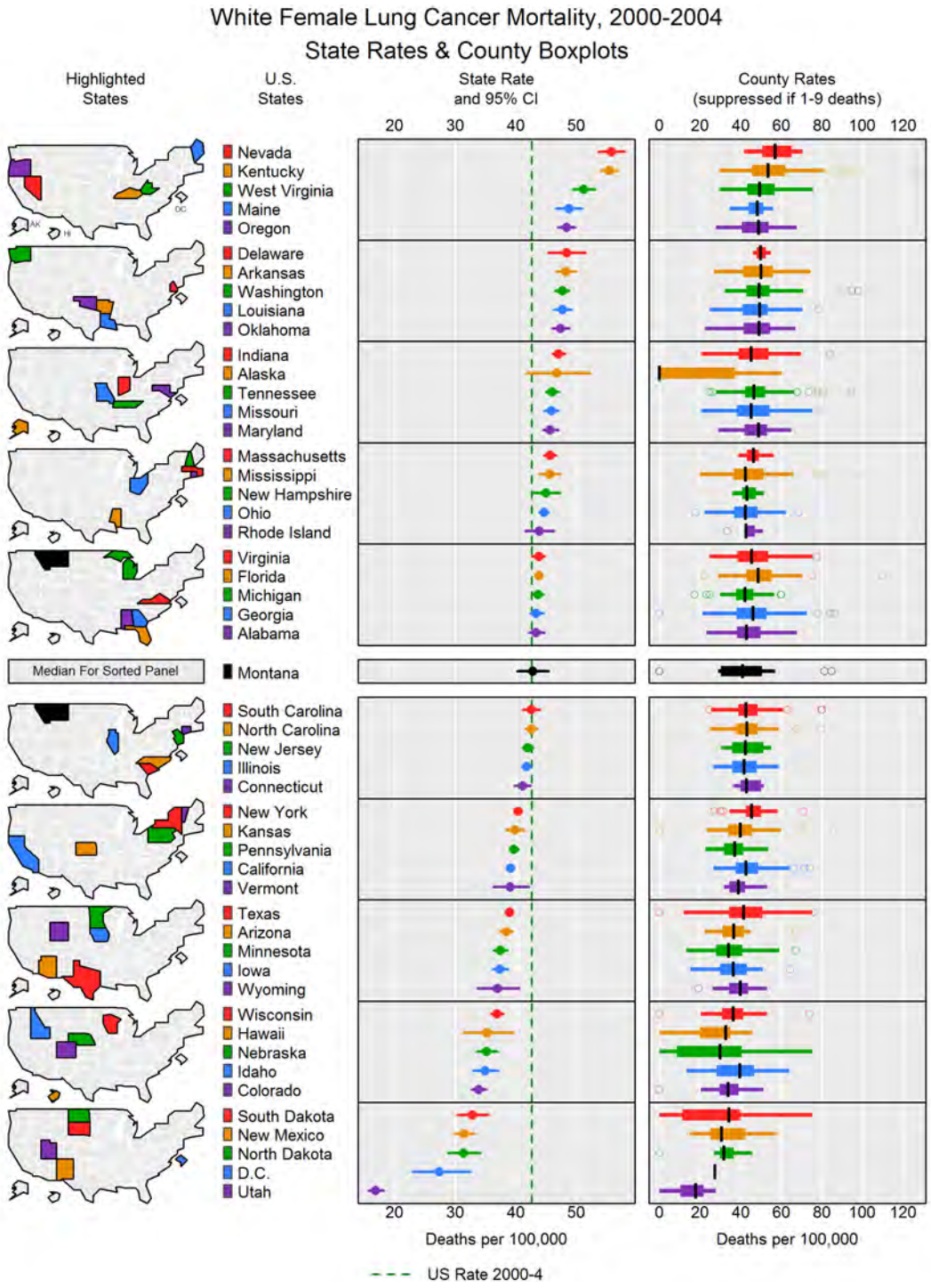

Figure 2: Dots with confidence interval lines and boxplots of county values within each state.

the type of column, such as a dot plot, where to get the data, such as column 3 of the input data frame (the first argument), and what text to use in the labeling lines. Each row in the plot description data frame describes a column in the plot. (This allows numeric descriptors

to go in numeric columns and text descriptors to go in text columns.) The R script example below shows entering values one column at a time using column labels and glyph type keywords that **micromapST** understands. An alternative to this description entry method is to create an empty data frame with `data.frame()` and then use `edit()` to fill in just the needed information. Note that this code does not produce any of the sample figures but simply lists all of the data frame elements.

```
R> panelDesc <- data.frame(type = c("map", "id", "arrow", "bar"),
+    lab1 = c("", "", "Rate", "Rate"),
+    lab2 = c("", "", "1950-69 to 1970-94", "1970-94"),
+    lab3 = c("", "", "Deaths per 100,000", "Deaths per 100,000"),
+    col1 = c(NA, NA, 1, 5), col2 = c(NA, NA, 3, NA),
+    panelData = c(NA, NA, NA, NA), refVals = c(NA, NA, NA, 100),
+    refTexts =  c(NA, NA, NA, "US Value"))
```

The `type` vector specifies the columns as a US state map on the left, followed by the state identifier (abbreviation, full name or fips code), then an arrow plot and a bar plot. The `lab1`, `lab2` and `lab3` vectors provide up to three rows of labels for the graphical columns (two above the column, one below); labels for the map and state i.d. columns are fixed internally and so are null values (`""`) in these lists. The `col1` and `col2` vectors provide column number indices or column names of the input data frame to be used by each glyph. `panelData` is shown for completeness but is only needed for more complex glyphs. Because all entries are `NA`, it could be dropped. `refVals` and `refTexts` provide an optional value (either numeric or a calculation) and label, respectively, for a vertical dashed green reference line in each column. Because the reference value and label are in the fourth position, the reference line will be drawn down the fourth column of panels, underneath all of the bars.

This example requests a simple US state map (`type = "map"` in `panelDesc`), but three types of cumulative shading can be requested. States that are highlighted in other panels are shaded light yellow and outlined in black, with the method of cumulative shading defined by the `type` option.These methods will be illustrated by individual examples in the next section but are combined in Figure 3 for ease of comparison. Cumulative shading from top to bottom (`mapcum`; Figure 3, second column) can be useful for spotting growing clusters of similar-valued states from top to bottom across the panels. `maptail` highlights states cumulatively from the top and bottom maps toward the median (Figure 3, third column). Another option is to use the light yellow shading to categorize the states into those that are above and below the median state (`mapmedian`; Figure 3, fourth column), displaying a dichotomy of high and low states in each map. These options can help the reader see changing geographic patterns across multiple maps or to find a state of interest. For example, a common visual query is "where is my state?" A glance at the top map with median shading quickly shows whether that state is in the top or bottom half of the ranked states, thus cutting search time in half.

Table 1 provides a complete list of glyphs available in **micromapST** and the corresponding information needed for each that must be passed to the program via the `col` or `panelData` parameters. Examples of each type of glyph will be given in the next section. A dot or bar needs only a single data value for plotting, so only `col1` needs to be included. Most of the other glyphs require two data values and so both `col1` and `col2` must be included. `col3` is only used for dots with confidence intervals. Box plots and the time series plots are more complex, so data for these glyphs are specified by `panelData`. The data values to be plotted

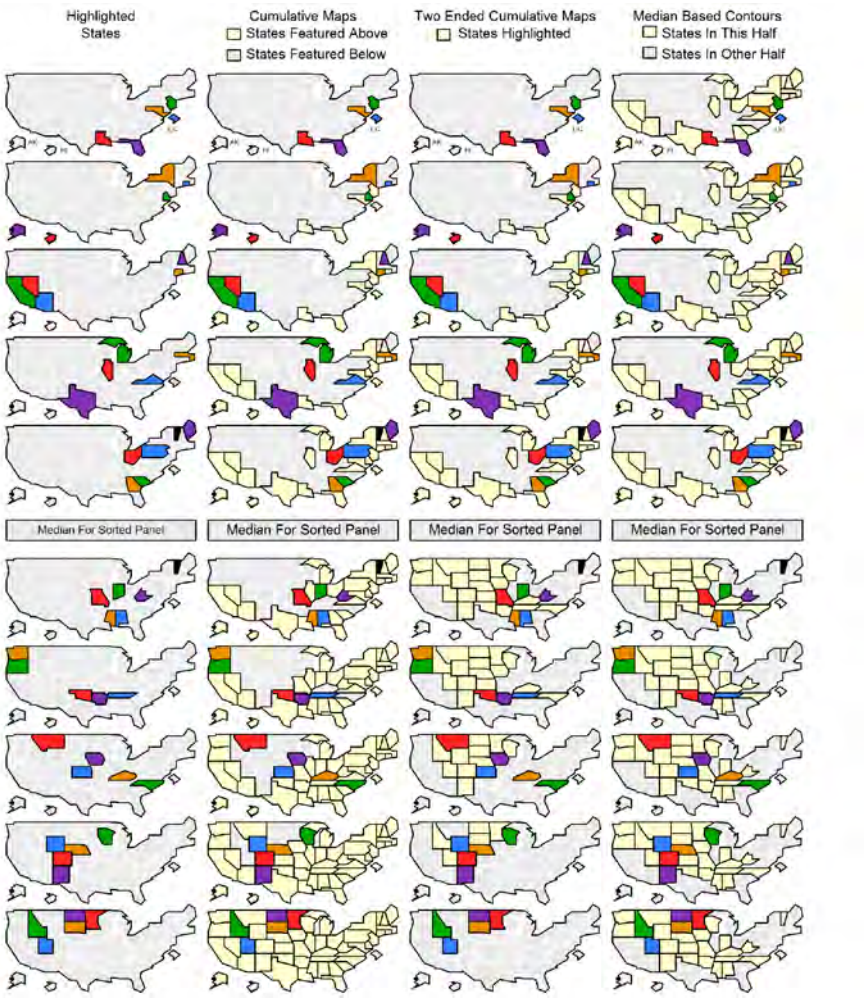

Figure 3: Types of cumulative map shading that can be requested. From left to right, the `type` in `panelDesc` equals `"map"`, `"mapcum"`, `"maptail"` and `"mapmedian"`.

are identified by their column numbers or names in the input data frame. For example, if our input data frame is named `mydata`, the `panelDesc` example above specifies that the third column of the linked micromap plot will consist of arrows with a starting value in `mydata[,1]` (third entry of `col1 = c(NA, NA, 1, 5)`) and an ending (arrowhead) value in `mydata[,3]` (`col2 = c(NA, NA, 3, NA)`). The fourth column of the linked micromap plot will be bars displaying the values in `mydata[,5]` (fourth entry of `col1 = c(NA, NA, 1, 5)`).

Once the panels and layout are defined, the function call provides the names of the input and panel description data frames, which must be the first two arguments in the call, respectively, and other details, with defaults shown below. For most applications, the defaults are acceptable, so that the function call is very short. Examples are given in the next section.

```
R> micromapST(mydata, panelDesc, rowNames = "ab", sortVar = NULL,
+    ascend = TRUE, title = c(" ", " "), plotNames = "full",
+    colors = stateMicromapDefaults$colors,
+    details = stateMicromapDefaults$details)
```

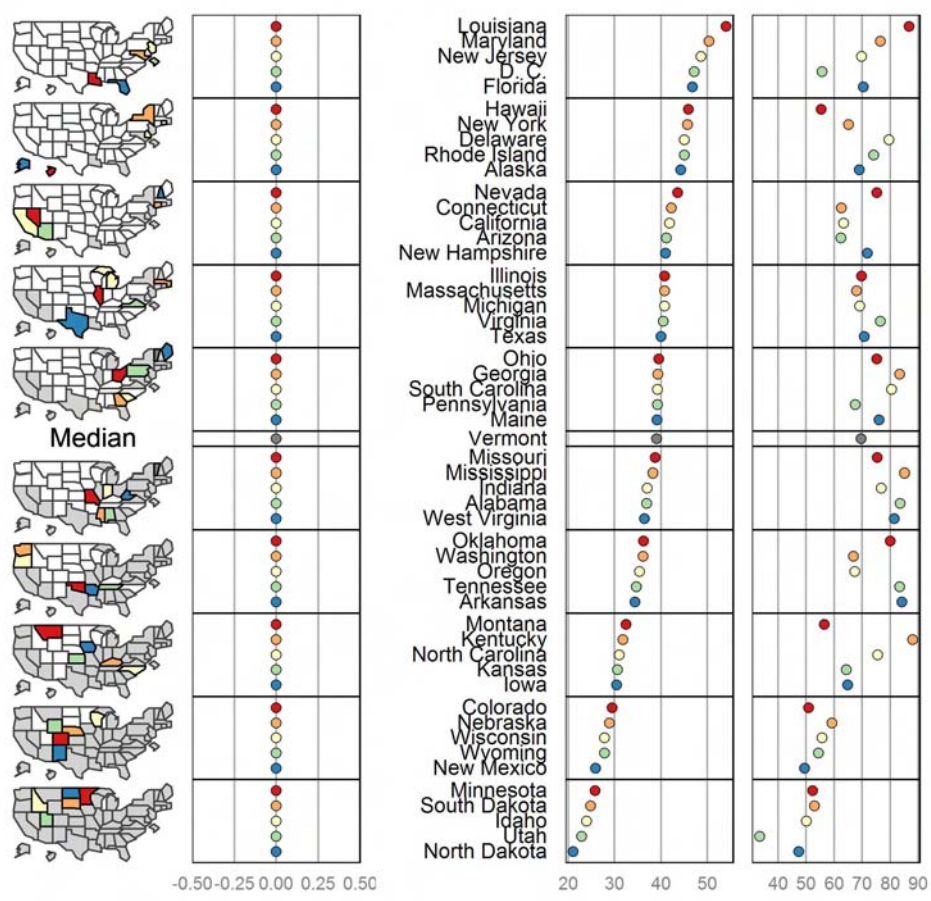

Figure 4: Two-column dot plot produced by the R package **micromap** using all default values.

Figure 5: Results of panel attribute modification to produce linking color squares and left justification of names in **micromap**.

The EPA package **micromap** can produce linked micromap plots with dots, bars and boxplots but requires more user-provided code to produce high-quality output. For example, Figure 4 is a two-column dot plot with maps, state names and a linking color dot produced with **micromap** (version 1.7) with all default settings:

```
R> lmplot(stat.data = wmlung5070, map.data = statePolys,
+    print.file = "Fig4-SquareswMedian all Defaults.jpeg",
+    panel.types = c("map", "dot", "labels", "dot", "dot"),
+    panel.data = list(NA, "points", "stName", "RATEWM_50", "RATEWM_70"),
+    ord.by = "RATEWM_50", rev.ord = TRUE, grouping = 5, median.row = TRUE,
+    map.link = c("StateAb", "ID"))
```

The resulting linked micromap plot is usable for data exploration but needs some formatting changes to be publication ready. Formatting is specified in a panel attribute list. For example, the following code, when included with the code above, will change the wide panels with linking color dots to a narrow column with colored squares and left-align the state names, similar to our design:

```
R> panel.att <- list(
+    list(2, panel.width = 0.11, right.margin = -0.25, point.type = 15,
+      point.size = 1.25, xaxis.text.display = FALSE,
+      xaxis.line.display = FALSE, graph.grid.major = FALSE,
+      graph.border.color = "white"),
+    list(3, panel.width = 0.8, align = "left", text.size = 0.9,
+      left.margin = -0.25)
+  )
```

This additional code seems simple enough, but the choice of parameter values, such as the panel and margin widths, must be determined by trial and error. This can be a tedious process for every change to the default design. A user could start with example code from the package's documentation, but changes would then be needed for the different geography, dataset format and plotted column specifications. It may be simpler to start with default values as we have done here or to mimic formatting options from **ggplot2** examples (Wickham 2009). There are also limitations in the **micromap** options available, such as the controls of label colors, the choices of output file formats and the requirement that full state names be included as a column in the data frame. In short, the **micromap** package is very flexible, but it can be time consuming to use because of the formatting modifications that must be made whenever the plot content changes. In the next section we will illustrate the types of glyphs that area available in **micromapST** along with the simple code required to produce them.

4. Examples of glyphs available in the micromapST package

4.1. Technical notes

Examples of the types of glyphs that can be included in a linked micromap plot using **micromapST** are described and illustrated in this section with accompanying code to produce them. The code for all examples in this section is compatible with R versions 2.15.3 and 3.0.3 and requires at least version 1.0.3 of **micromapST**. This package has been run successfully on Mac, Unix and Windows platforms and will send the resulting graphics image to any device that the user can specify in R. However, it has been our experience that some default settings for graphics devices may need to be fine tuned in order to provide sufficient resolution and size to display a high quality linked micromap image. For example, a default-size graphics window in R may be too small on your monitor, causing overplotting of the micromap panels. We have found that adding `windows(7.5, 10, xpinch = 72, ypinch = 72, point = 9)` to the R program prior to the first call to **micromapST** solves this problem on our own Windows system. The corresponding graphics window specification for a Mac OS X system is `quartz(width=7.5, height = 10, dpi = 72, pointsize = 10)`. Different monitors have different pixels per inch so that these parameters might need to be modified to produce the

Glyph name	Meaning	col1	col2	col3	panelData
arrow	Arrow	Beginning values	Ending values (arrowhead)	NA	NA
bar	Horizontal bar	Bar end values	NA	NA	NA
segbar	Horizontal stacked bar	Values for first (leftmost) bar segment	Values for last (rightmost) bar segment	NA	NA
normbar	Horizontal stacked bar, normalized to total 100%	Same as segbar	Same as segbar	NA	NA
ctrbar	Horizontal stacked bar, centered on middle bar	Same as segbar	Same as segbar	NA	NA
boxplot	Horizontal box plot	NA	NA	NA	Name of output list from call to boxplot(..., plot = FALSE)
dot	Dot	Values for dots	NA	NA	NA
dotconf	Dot with confidence interval line	Values for dots	Values of lower limits	Values of upper limits	NA
dotse	Dot with line lentgh of +/- standard error	Values for dots	Standard errors	NA	NA
scatdot	Scatter plot of dots	Values on horizontal (x) axis	Values on vertical (y) axis	NA	NA
ts	Time series (line) plot	NA	NA	NA	Name of array with dimensions c(51, t, 2) where t = # of time points, x values in [,,1], y values in [,,2]
tsconf	Time series (line) plot with confidence band	NA	NA	NA	Name of array with dimensions c(51, t, 4) as specified for ts with added lower limit in [,,3], upper limit in [,,4]

Table 1: Glyphs available in **micromapST** and the specification of input data by column name or number, which indexes the input data frame (NA indicates that this column parameter is not used for the glyph).

desired image size. For printer output, one of the platform-independent file formats can be used. For example, the sample code in this section includes `pdf(file = filename, width = 7.5, height = 10)` to show how to write the linked micromap image to a PDF file. In order to produce a clear image in LaTeX small enough for this journal, we produced 300 dpi PNG

files by using `png(file = filename, width = 7.5, height = 10, units = "in", res = 300)` and then scaled the images to about 90% of their original size. Similar parameters can be specified for TIFF and JPEG output.

4.2. Bars, dots and boxplots

A single statistical value may be represented by a bar or by a dot with or without error bars or confidence intervals. Figure 2 displays state white female lung cancer age-adjusted (2000 US standard) mortality rates aggregated for 2000-4 by dots with 95% confidence interval lines (`type = "dotconf"`), with the rows sorted in descending order by the state rates (Surveillance, Epidemiology, and End Results (SEER) Program, National Cancer Institute 2008). Although Nevada had the highest aggregated rate, its confidence interval totally overlaps that of Kentucky, so we can conclude that Nevada and Kentucky are not statistically different, but that these two states have rates significantly higher than the third ranking state, West Virginia. In the second column of panels, box plots (`type = "boxplot"`) display the distribution of county values within each state. We can see that some states have more variability across their counties than others. For example, Florida has a state rate near the US median rate (vertical dashed line) but has a county with a rate that is second highest of all displayed counties. (Data suppression rules required counties with 1 to 9 deaths to be omitted from the plot.) The **micromapST** code to produce this example and send the image to a PDF file is as follows:

```
R> data("wflung00and95", "wflung00and95US", "wflung00cnty")
R> wfboxlist <- boxplot(split(wflung00cnty$rate, wflung00cnty$stabr),
+    plot = FALSE)
R> panelDesc <- data.frame(type = c("map", "id", "dotconf", "boxplot"),
+    lab1 = c("", "", "State Rate", "County Rates"),
+    lab2 = c("", "", "and 95% CI", "(suppressed if 1-9 deaths)"),
+    lab3 = c("", "", "Deaths per 100,000", "Deaths per 100,000"),
+    col1 = c(NA, NA, 1,NA), col2 = c(NA, NA, 3, NA),
+    col3 = c( NA, NA, 4, NA), refVals = c(NA, NA, wflung00and95US[1,1], NA),
+    refTexts = c(NA, NA, "US Rate 2000-4", NA),
+    panelData = c("", "", "", "wfboxlist"))
R> pdf(file = "WFLung-2000-2004-State-Dot-County-Box.pdf",
+    width = 7.5, height = 10)
R> micromapST(wflung00and95, panelDesc, sortVar = 1, ascend = FALSE,
+    title = c("White Female Lung Cancer Mortality, 2000-2004",
+      "State Rates & County Boxplots"))
R> dev.off()
```

4.3. Arrows

The simplest measure of change between two values is the percent increase or decrease, but this omits the context of magnitudes of the values. Two values may be represented by a directional arrow, allowing the reader to judge the magnitude of change by the length of the arrow and to estimate the two values by the position of the arrow endpoints. Figure 6 displays the change in white male lung cancer age-adjusted mortality rates from the aggregated period 1950–69 to

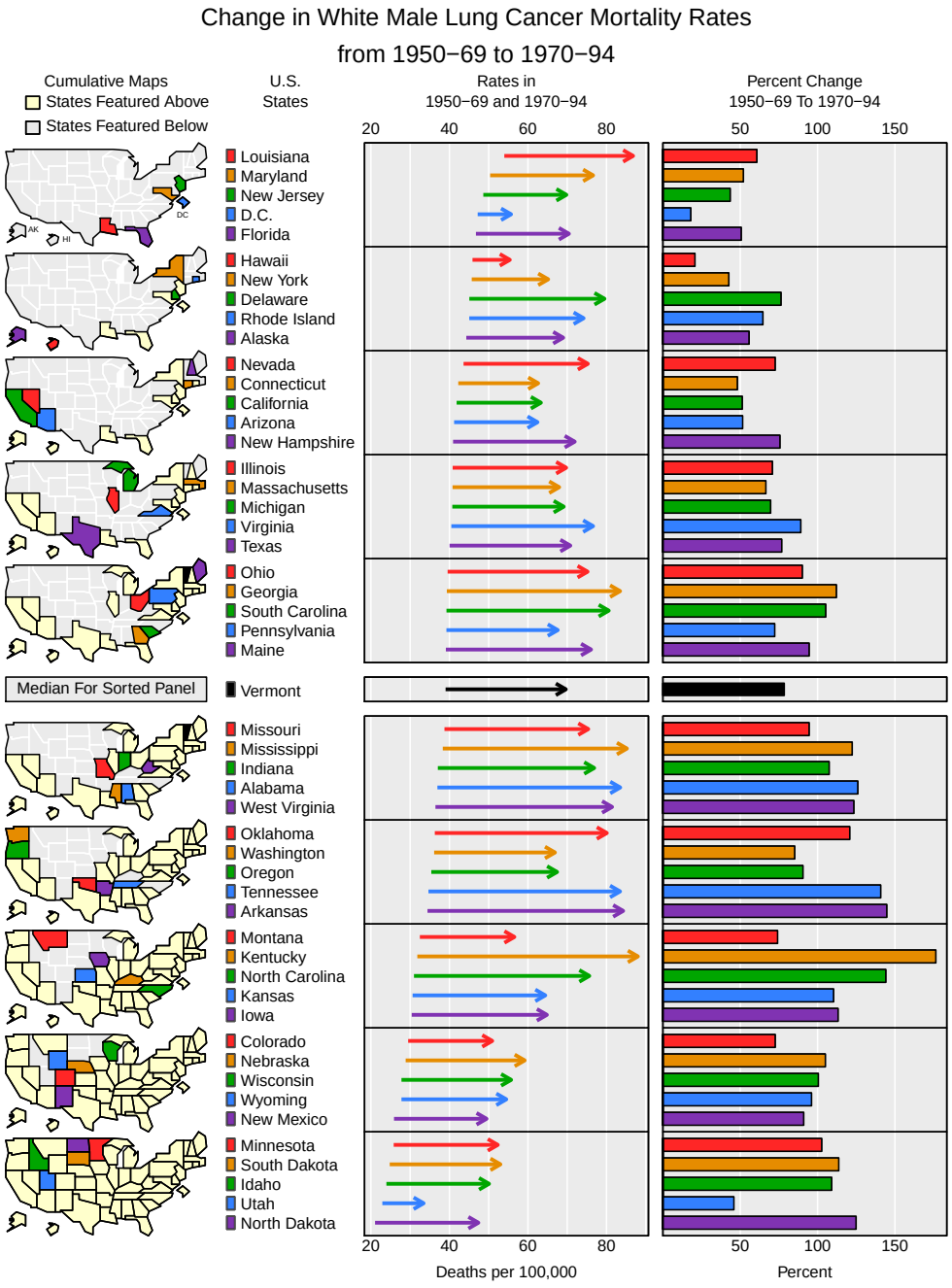

Figure 6: Change in white male lung cancer mortality rates from 1950–74 to 1975–94 represented as arrows (range of rates) and bars (percent change), sorted by rates in 1950–69.

1970–94 by both of these measures (`type = "arrow"` and `type = "bar"`) (Devesa, Grauman, Blot, Pennello, Hoover, and Fraumeni Jr. 2005). This design facilitates three different tasks: (1) comparison of 1950–69 rates by focusing on the left end of the arrows (the sort variable); (2) comparison of absolute rates of change between time periods by focusing on arrow lengths; and (3) comparing the percent change of the rates by scanning down the bars in the second column. The **micromapST** code to produce this example and create a PDF file:

```
R> data("wmlung5070", "wmlung5070US")
R> panelDesc <- data.frame(type = c("mapcum", "id", "arrow", "bar"),
+    lab1 = c("","", "Rates in", "Percent Change"),
+    lab2 = c("","", "1950-69 and 1970-94", "1950-69 To 1970-94"),
+    lab3 = c("","", "Deaths per 100,000","Percent"),
+    col1 = c(NA, NA, 1, 5), col2 = c(NA, NA, 3, NA))
R> pdf(file = "WMLung50-70-Arrow-Bar.pdf", width = 7.5, height = 10)
R> micromapST(wmlung5070, panelDesc, sortVar = 1, ascend = FALSE,
+    title = c("Change in White Male Lung Cancer Mortality Rates",
+      "from 1950-69 to 1970-94"))
R> dev.off()
```

4.4. Time series plots

For the dot, bar, arrow and box plot, each state is represented by a single row of glyphs in its panel. However, there are other types of displays for which each panel serves as a two-dimensional graph, including glyphs for all 5 states in that panel. For example, Figure 7 shows the time series of female lung cancer age-adjusted mortality rates annually from 1996 to 2010 (Surveillance, Epidemiology, and End Results (SEER) Program, National Cancer Institute 2013). The input data frame consists of a year label, the point value and the low and high limit values for each year/state combination. The line color in the plot still identifies the state but since the entire panel is the graph area, state lines can cross. Each state dot is plotted over a confidence band of the same color (`type = "tsconf"`), constructed by connecting the 95% confidence interval endpoints for each state across the 15 years. The large green confidence interval for Alaska in the second panel consumes most of the vertical resolution available; in fact, the common y axis limits of all the panels are determined by the high confidence limits for Alaska and the low values for Utah. Despite the overplotting, we can see some patterns in the time series that would not be apparent in a simple bar or dot plot. In this example, we can see that rates have converged over time for states in all but the highest and lowest panels. It is also clear that Kentucky rates have been high and Utah rates have been low over the entire period and Nevada rates have dropped sharply since 1996. The **micromapST** code to produce this example and send the image to a PDF file is as follows:

```
R> data("TSdata")
R> temprates <- data.frame(TSdata[,,2])
R> panelDesc <- data.frame(type = c("maptail", "id", "tsconf", "dot"),
+    lab1 = c("", "", "Time Series", "Female"),
+    lab2 = c("", "", "Annual Rate per 100,000", "Most Recent Rate (2010)"),
+    lab3 = c("", "", "Years", "Deaths per 100,000"),
+    lab4 = c("", "", "Rate", ""), col1 = c(NA, NA, NA, 15),
+    panelData = c(NA, NA, "TSdata", NA))
R> ExTitle <- c("Time Series with Confidence bands",
+    "Annual Female Lung Cancer Mortality Rates, 1996-2010")
R> pdf(file = "Time-Series-with-Conf.pdf", width = 7.5, height = 10)
R> micromapST(temprates, panelDesc, sortVar = 15, ascend = FALSE,
+    title = ExTitle)
R> dev.off()
```

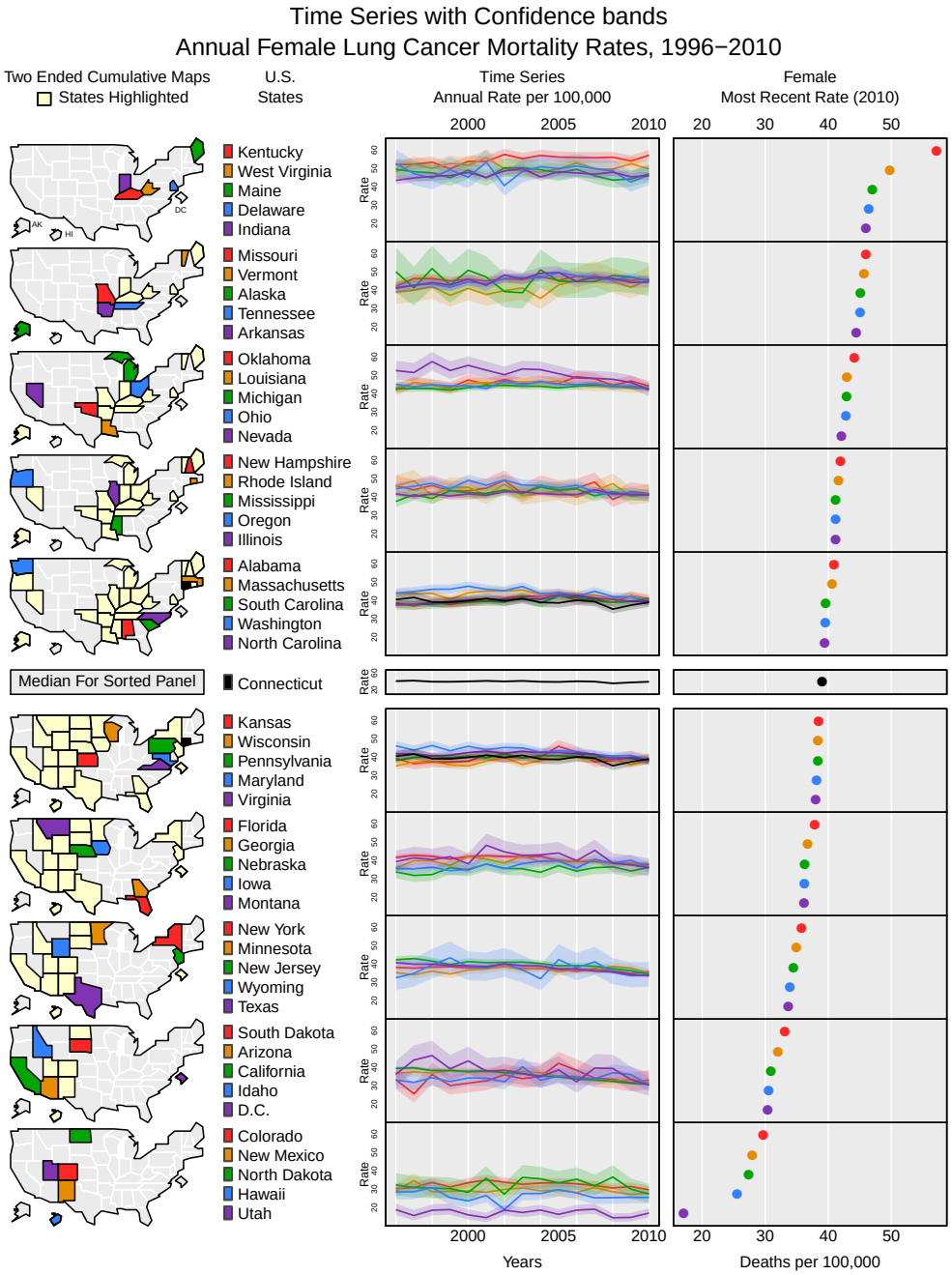

Figure 7: Time series plots of female lung cancer mortality rates from 1996 to 2010 with 95% confidence interval bands, sorted in descending order by the most recent rate.

Cleveland (1993) recommended that the aspect ratio for time series graphs be chosen so that the average of the many line segments is about 45 degrees, the angle perceived accurately by most viewers. Because the **micromapST** page design has 10 small maps, the vertical space is fixed for each panel. The horizontal space, however, is dependent on the number of statistical columns requested – more columns result in narrower panels, increasing each graph's aspect ratio. Since Cleveland's recommendation is data dependent, adherence to it would require

specifying panel widths according to the time series' data values, something not possible in the current package. Dropping confidence intervals in Figure 7 would increase vertical resolution and remove the color blending of overlapping interval bands. This one-page design with its fixed panel height means that tasks requiring more y axis resolution cannot be well served.

There are alternative approaches to showing time series, such as the sparkline plot suggested by Tufte (2006). Conversely, the linked micromap time series plot can be used for other types of bivariate data. For example, if the x and y values represent national and state percentiles of the data distribution, respectively, then the line plot would provide a comparison of the two distributions, similar to a Q-Q plot.

4.5. Scatter plots

Another method of comparing values is to show where each state's pairwise values fall relative to the distribution of bivariate values for all of the states. For example, each panel in the second panel column of Figure 8 represents a scatterplot (`type = "scatdot"`) of the endpoints of the time series data (1996 on horizontal axis and 2010 on vertical axis) used for Figure 7. The dots for the five states in a particular row of panels are filled with the colors representing those five states. The dots for the other 46 states are shown to provide context for these five states relative to the full distribution of 51 pairs of values. Overplotting can be reduced by using open circles, as in Figure 8, by reducing their size or by an algorithmic shifting of overlapping dots. The point for the median state, based on the sort variable, is represented as a black-filled dot. The background dots are printed first, then the median and finally the color-filled dots so that the most important information for that panel is visually in the foreground. As noted above for time series, the vertical space constraints force the scatter plot to fit into a small rectangle much wider than it is high. To compensate for what may be a sub-optimal aspect ratio for many data values, a white diagonal line is drawn indicating equality of the x and y values. In this example, nearly all of the state dots are below this diagonal line, indicating that rates had declined from 1996 to 2010 in nearly all states. Consistent with the conclusions from Figure 7, it is also clear that Kentucky and Utah were at the extremes of the state rate range in both 1996 and 2010, with their dots falling in the upper right of the highest panel and lower left of the lowest panel, respectively. The decline in Nevada's rates is also clear in the top panel, where its dot is farther from the equality line than those of the other four states in the top group. The **micromapST** code to produce this example is:

```
R> panelDesc <- data.frame(type = c("mapmedian", "id", "dot", "scatdot"),
+    lab1 = c("", "", "Female Lung Cancer Mortality", "Comparison of Rates"),
+    lab2 = c("", "", "Rate in 1996 (Sort Variable)",
+      "in 1996 (x axis) and 2010 (y axis)"),
+    lab3 = c("", "", "Deaths per 100,000", "Deaths per 100,000 in 1996"),
+    lab4 = c("", "", "", "Rate in 2010"), col1 = c(NA, NA, 1, 1),
+    col2 = c(NA, NA, NA, 15))
R> ExTitle <- c("Dot Plot for 1996, Scatter Plot Comparing 1996 to 2010",
+    "Female Lung Cancer Mortality Rates")
R> pdf(file = "Scatter-Dots.pdf",width=7.5,height=10)
R> micromapST(temprates, panelDesc, sortVar = 1, ascend = FALSE,
+    title = ExTitle)
R> dev.off()
```

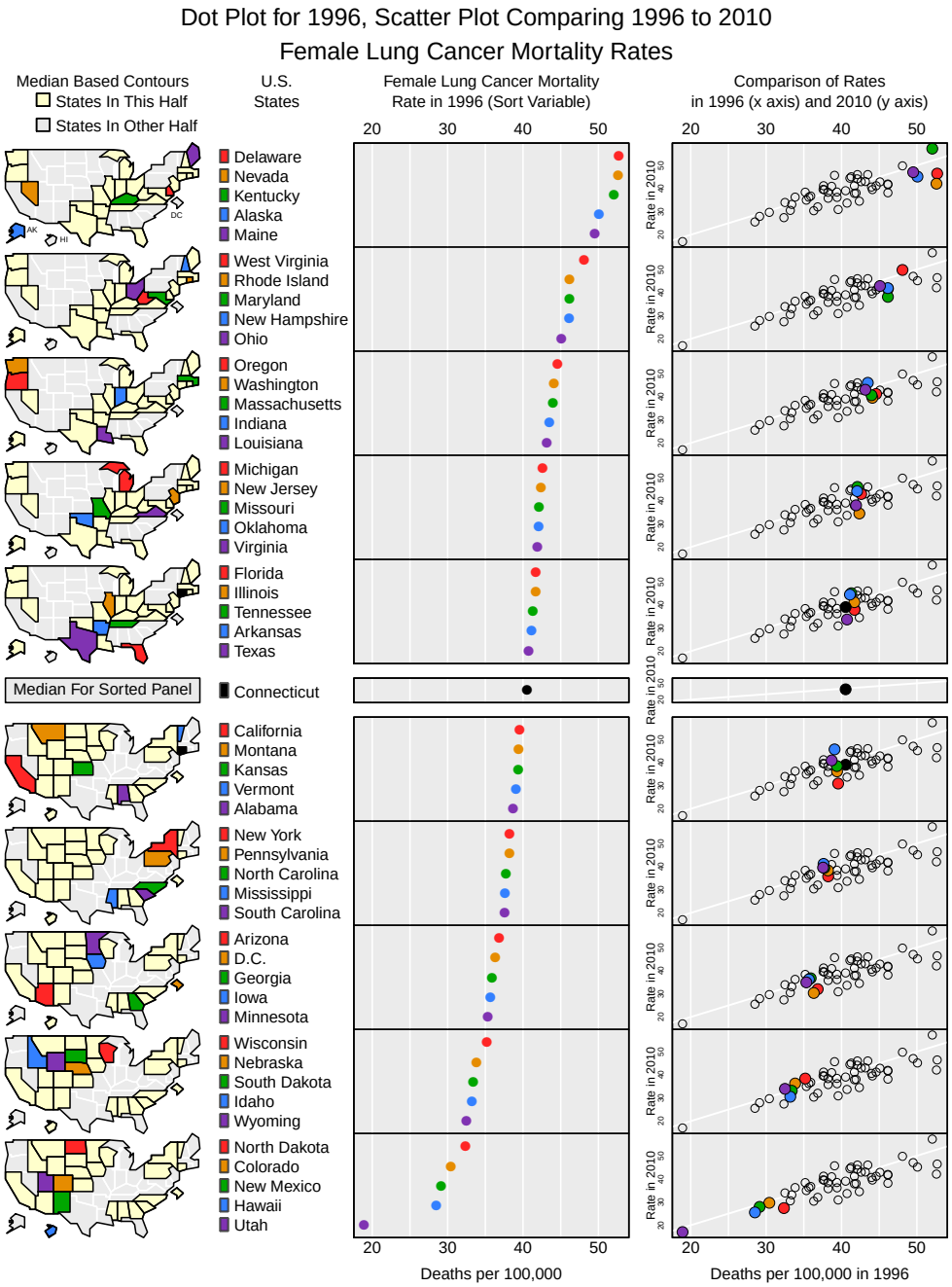

Figure 8: Scatter plot displaying 1996 vs. 2010 female lung cancer mortality rates.

4.6. Alternative scatter plot designs

The design for scatter plots in linked micromaps will be adequate for most applications, but occasionally more vertical plotting space is needed to reduce overplotting and improve the aspect ratio as noted above, e.g., for Q-Q plots and time series data. The specialized design for US states presented in this paper can be adapted to double the vertical axis space by using two designs described by Carr (Carr 2001; Carr and Pickle 2010). Both are two-column designs

that use a combination of symbol shapes and colors to display more than the 5 geographic units per perceptual group in the **micromapST** design. Because shape and color are separable encodings, we can tune our vision to find a distinct shape and then a color, a task made easier when the sort order localizes the squares (or colors) in the scatter plot.

We used these same ideas to create a square scatter plot design for all US states. This option is not implemented in **micromapST**, but is simply a prototype. This design, shown in Figure 9, includes 10 scatter plots with 5 or 6 states highlighted in each. However, the plots are twice as high as in Figure 8 and so are arranged in two columns, requiring the reader to alternate plot reading from left to right. That is, the first 5 states of the ranked group of 10 are plotted as circles on the leftmost plot and the second 5 states are plotted as squares on the rightmost plot. The advantage of this design is the greater vertical plot resolution, but the cost is a longer visual search required to mentally link between the two plots and the state name or map, which is farther from the rightmost plot. Testing is required to determine whether its advantages outweigh the disadvantages before adding it to our package.

Figure 9 presents scatterplots of male versus female lung cancer mortality rates for 2010. The states are sorted in ascending order by a computed, not plotted, value, i.e., the relative difference between males and females. Although Kentucky had the highest rates of all states for both men and women in 2010, the male and female rates in 35 other states are closer to each other, with the smallest differences seen in Alaska and Nevada at the top of the plots. Similarly, Utah has the lowest female rate but Mississippi has the largest relative difference between the male and female rates (bottom of the plots). The scatterplot suggests a linear relationship between the male and female rates. The ratio used to sort the states tends to highlight points along slices parallel to this suggested line and to move progressively from one side to the other. Since the units of measure for the male and female rates are the same, alternatives for assessing discrepancies between the plotted variables include orthogonal or principal curves regression (Hastie and Stuetzle 1989) and comparison of signed distances (Zhang 2012).

4.7. Bar charts for categorical data

Another type of data that can be displayed easily is categorical data, displayed by **micromapST** as horizontal stacked bar charts. The audience for this design is anyone who needs to display rating scale data, population pyramids or survey results reported as counts or percents, e.g., on Likert scale (Likert 1932). The use of color as a link throughout a perceptual group is not compatible with using a distinct color for each category within the state bar, so a light to dark sequence of the same color is used to represent the values in each bar segment corrresponding to the data categories.

Cleveland and McGill (1984) reported that we can accurately judge length along bars or lines with a common starting point. The problem of judging length along non-aligned scales can be at least partially alleviated by centering the bars on the middle of the categories for each state, as suggested by Heiberger and Robbins (Heiberger and Robbins 2014) and implemented in the R package **HH** (Heiberger 2014), converting it to a diverging bar chart. This design can be used to highlight positive and negative categories, such as survey results reported on a Likert scale. The **micromapST** package will display up to 9 categories and will center the bars at the middle of the middle bar if there is an odd number of categories, and at the boundary of a bar segment that will divide an even number of bars equally (`type`

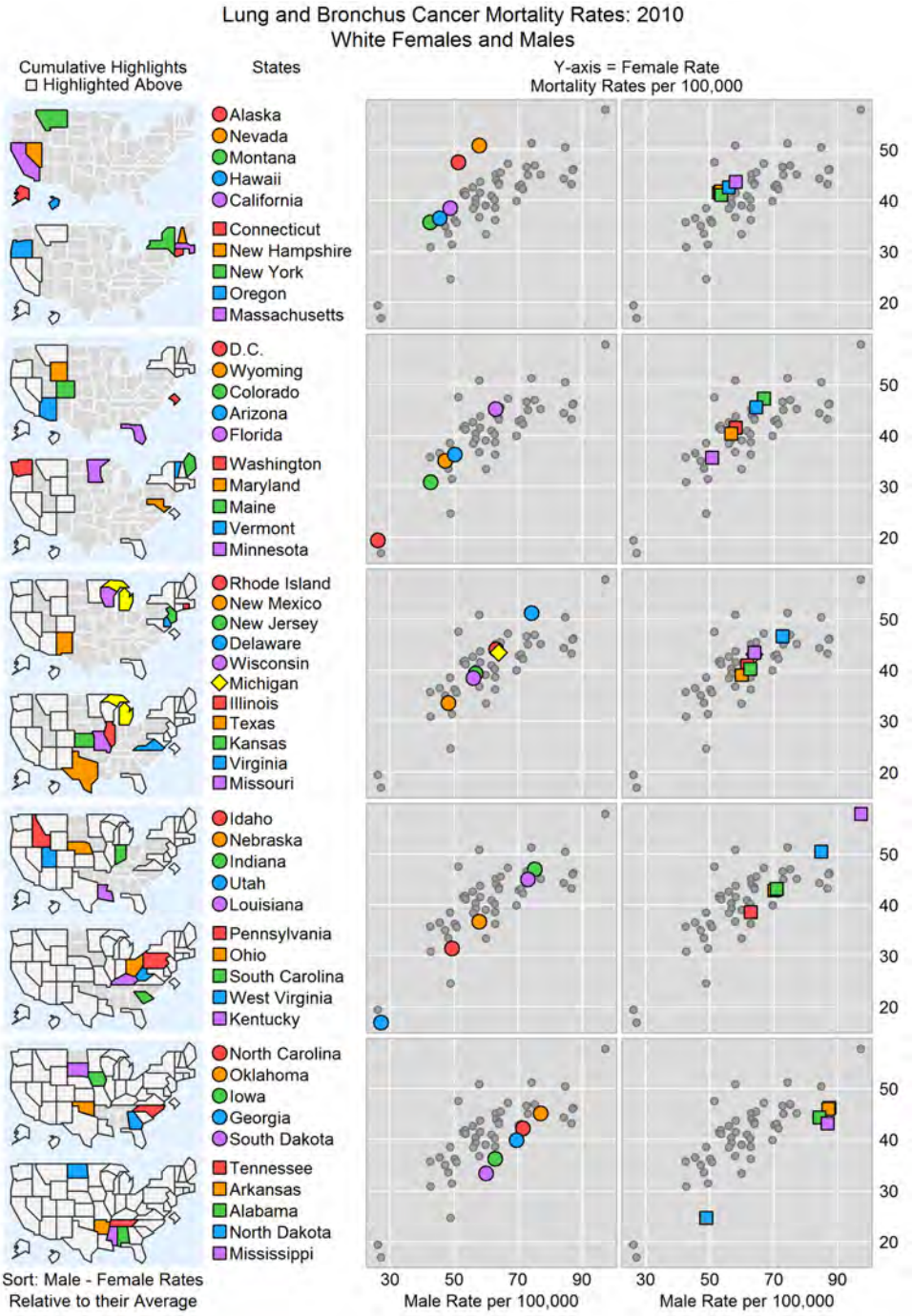

Figure 9: Alternative scatter plot design with states sorted by the relative difference between male and female rates ((male rate − female rate)/(average of male and female rates)).

`= "ctrbar"`). Other options available are bars that are left justified (`type = "segbar"`) or normalized to a length of 100% (`type = "normbar"`). Figure 10 uses this design to display the percents of eighth grade students with math proficiency scores in four categories: less than basic, basic, proficient, and advanced (National Center for Education Statistics 2013).

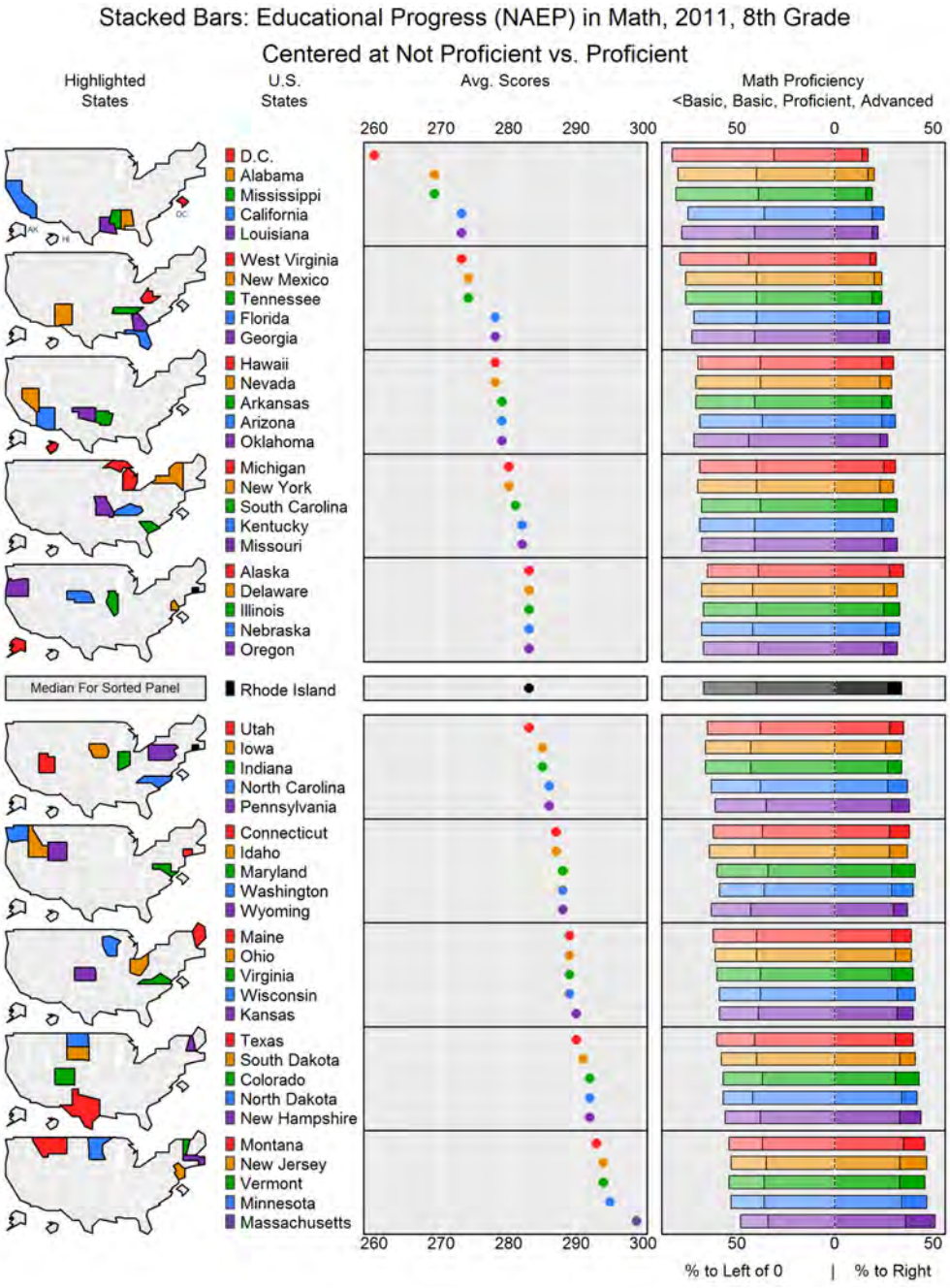

Figure 10: Centered (diverging) horizontal stacked bars.

The state bars are centered between the second and third of the four categories, so that the total length of the bar segments to the left and right of center can be interpreted as the percent of students who were less than proficient and at least proficient, respectively. One interesting observation from this display is that the District of Columbia had a much lower average score (the sorting variable), than the second lowest state, Alabama. However, the percents of students who were less than proficient (represented by the length of the bar to the left of zero) are not that different; differences between these two states is primarily between

the basic and less than basic categories (represented by the lengths of the two leftmost bar segments). The **micromapST** code to produce this example and send the image to a PDF file is as follows:

```
R> data("Educ8thData")
R> panelDesc <- data.frame(type = c("map", "id", "dot", "ctrbar"),
+    lab1 = c("", "", "Avg. Scores", "Math Proficiency"),
+    lab2 = c("", "", "", "<Basic, Basic, Proficient, Advanced"),
+    lab3 = c("", "", "", "% to Left of 0           |    % to Right"),
+    col1 = c(NA, NA, 3, 4), col2 = c( NA, NA, NA, 7))
R> ExTitle <- c(
+    "Stacked Bars: Educational Progress (NAEP) in Math, 2011, 8th Grade",
+    "Centered at Not Proficient vs. Proficient")
R> pdf(file = "Educ-Centered-Bar.pdf", width = 7.5, height = 10)
R> micromapST(Educ8thData, panelDesc, sortVar = 3, title = ExTitle)
R> dev.off()
```

5. Discussion

The linked micromap plot design was initially created for EPA applications (Carr *et al.* 1998) and has been used by several federal statistical agencies such the National Agricultural Statistics Service and the National Cancer Institute. NCI's web-based implementation of linked micromaps has been a popular and effective way to communicate cancer statistics, even to a non-statistical audience (Bell *et al.* 2006). Since 2011, an early version of our **micromapST** package has been used in several statistical graphics classes at George Mason University (GMU), including one undergraduate class. The students readily produced the assigned graphics and some made use of the package in their data exploration projects and poor graph redesign. Adding spatial context to state data often provided an improvement to the original graphs. It was easy for them to find state data for these projects and to use this package, freeing them to focus on the statistical and visualization tasks.

The diversity of students at GMU led to projects involving other nations of world, such as Peru and Ghana. The web site for global administrative areas (`http://www.gadm.org/`) has Level 1 and Level 2 shapefiles for administrative units analogous to US states and counties. This source can serve as starting points for development of suitable Level 1 visibility maps for other countries, with simplified boundaries and modified polygons areas for enhanced color identification. The evolution of **micromapST** is likely to support Level 1 visualization for an increasing number of nations in the future.

While the linked micromap design has been positively received, its spread has been substantially slowed by the lack of convenient software tools for plot production. Our package **micromapST** and the EPA package **micromap** are part of an effort to make linked micromaps more widely available. The two packages spring from the same origins and exhibit most of the same design considerations. The linked micromap is a complex design, so specifying all of the details is not straightforward. By focusing on a narrow range of tasks, such as comparing US states, the specification can be simplified. The page layout can be standardized for the particular task, requiring the user only to provide and define the plot content. This is the approach we have taken in the development of **micromapST**.

The EPA **micromap** package was developed for EPA applications but its generality makes it relevant for many other types of data. Its designs follow many of the same criteria as our work and can produce excellent linked micromap plots. However, fine tuning some of the designs may take work. The **ggplot2** package that underlies the **micromap** package provides great flexibility but can be noticeably slow in running and its syntax may be a deterrent to those who are not familiar with the finer details of **ggplot2**. Because **micromap** is also an R package, the developers and sophisticated users should, with sufficient effort, be able to emulate **micromapsST** graphics as new design variations emerge.

The current **micromapST** package focuses exclusively on showing US states, with the ability to show the distribution of counties within each state through boxplots. The most recent version adds time series confidence bands, scatter plots and stacked bar charts, the latter in order to attract more of the social science community. Future enhancements include additional types of graphics and improvements to the new stacked bar design, such as a different color scheme; application to other geographic units beyond US states; refinement of labeling and white space calculation, modified panel widths and separations. We would like to simplify the input of plot details and plan to provide a GUI interface. The popularity of the NCI interactive, web-based version of linked micromaps is encouraging and we are exploring options to implement a web-based version, such as by using **shiny** (RStudio, Inc. 2014).

The new **micromapST** package allows an analyst with state-level data to quickly explore and communicate that data using the linked micromap plot design. By basing the design on principles of perception and cognition, the linked micromap can enhance the amount of information the reader obtains from the visualization. The ease of use should encourage interactive exploration for state data. Minimal R coding is needed to prepare state data for use with this package and the layout of the resulting plot is usually of sufficient quality for publication. We think the narrower focus of **micromapST** will make it easier for students and those who are not interested in the details of R or **ggplot2** syntax to learn and use the package and for developers to enhance it. All designs involve compromise. Our primary goal of displaying all 51 states on a single page constrains the vertical space for the small graphics, which may not be optimal for all types of data. However, to quote Tufte (Tufte 1983, p. 51), "Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space". We think that the linked micromap design meets this objective.

References

Bell BS, Hoskins RE, Pickle LW, Wartenberg D (2006). "Current Practices in Spatial Analysis of Cancer Data: Mapping Health Statistics to Inform Policymakers and the Public." *International Journal of Health Geographics*, **5**, 49.

Bertin J (1973). *Semiologie Graphique*. Mouton-Gautier, The Hague.

Blunt G (2006). "Using **grid** Graphics to Produce Linked Micromap Plots of Large Financial Datasets." In *useR! 2006*. Vienna. URL http://www.R-project.org/user-2006/Abstracts/.

Carr DB (1997). "Some Simple S-PLUS Tools for Matrix Layouts." *Technical report*, Bureau of Labor Statistics. Statistical Note Series.

Carr DB (2001). "Designing Linked Micromap Plots for States with Many Counties." *Statistics in Medicine*, **20**, 1331–1339.

Carr DB, Olsen AR (1996). "Simplifying Visual Appearance by Sorting: An Example Using 159 AVHRR Classes." *Statistical Computing and Graphics Newsletter*, **7**, 10–16.

Carr DB, Olsen AR, Courbois JP, Pierson S, Carr DA (1998). "Linked Micromap Plots: Named and Described." *Statistical Computing and Graphics Newsletter*, **9**, 24–32.

Carr DB, Pearson Jr JB, Pickle LW (2013). ***micromapST**: State Linked Micromap Plots*. R package version 1.02, URL http://CRAN.R-project.org/package=micromapST.

Carr DB, Pickle LW (2010). *Visualizing Data Patterns with Micromaps*. Chapman & Hall/CRC, Boca Raton.

Carr DB, Pierson SM (1996). "Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps." *Statistical Computing and Graphics Newsletter*, **7**, 16–23.

Carr DB, Wallin JF, Carr DA (2000). "Two New Templates for Epidemiology Applications: Linked Micromap Plots and Conditioned Choropleth Maps." *Statistics in Medicine*, **19**, 2521–2538.

Cleveland WS (1993). *Visualizing Data*. Hobart Press, Summit.

Cleveland WS, McGill R (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association*, **79**, 531–554.

Devesa SS, Grauman DJ, Blot WJ, Pennello GA, Hoover RN, Fraumeni Jr JF (2005). *Atlas of Cancer Mortality in the United States, 1950–1994. NIH Publication No. 99-4564, 1999*. NIH.

Friendly M, Kwan E (2003). "Effect Ordering for Data Displays." *Computational Statistics & Data Analysis*, **43**, 509–539.

Gebreab SY, Gillies RR, Munger RG, Symanzik J (2008). "Visualization and Interpretation of Birth Defects Data Using Linked Micromap Plots." *Birth Defects Research Part A – Clinical and Molecular Teratology*, **82**, 110–119.

Hastie T, Stuetzle W (1989). "Principal Curves." *Journal of the American Statistical Association*, **84**, 502–516.

Heiberger RM (2014). ***HH***. R package version 3.0-3, URL http://CRAN.R-project.org/package=HH.

Heiberger RM, Robbins NB (2014). "Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications." *Journal of Statistical Software*, **57**(5), 1–32. URL http://www.jstatsoft.org/v57/i05/.

Kosslyn SM (2006). *Graph Design for the Eye and Mind*. Oxford University Press, New York.

Likert R (1932). "A Technique for the Measurement of Attitudes." *Archives of Psychology*, **140**, 1–55.

Monmonier M (1988). "Geographical Representation in Statistical Graphics: A Conceptual Framework." In *Proceedings of the Section on Statistical Graphics 1988*, pp. 1–10.

Monmonier M (1993). *Mapping It Out: Expository Cartography for the Humanities and Social Sciences*. University of Chicago Press, Chicago.

National Cancer Institute (2013a). "NCI Tools, Linked Micromaps." URL http://gis.cancer.gov/tools/micromaps/.

National Cancer Institute (2013b). "State Cancer Profiles Linked Micromaps." URL http://statecancerprofiles.cancer.gov/micromaps/.

National Center for Education Statistics (2013). "National Assessment of Educational Progress (NAEP), Grade 8, Mathematics, 2011." URL http://nces.ed.gov/nationsreportcard/states/.

Payton Q, McManus M, Weber M, Olsen T, Kincaid T (2015). "**micromap**: A Package for Linked Micromaps." *Journal of Statistical Software*, **63**, 1–16. URL http://www.jstatsoft.org/v63/i02/.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

RStudio, Inc (2014). ***shiny**: State Linked Micromap Plots*. R package version 0.10.2.1, URL http://CRAN.R-project.org/package=shiny.

Surveillance, Epidemiology, and End Results (SEER) Program, National Cancer Institute (2008). "SEER*Stat software, November 2007 data submission." Data originally provided to NCI by the National Center for Health Statistics, URL http://www.seer.cancer.gov/seerstat.

Surveillance, Epidemiology, and End Results (SEER) Program, National Cancer Institute (2013). "SEER*Stat Database: Mortality – All COD, Aggregated With State, Total U.S. (1969-2010) (Katrina/Rita Population Adjustment)." Data originally provided to NCI by the National Center for Health Statistics, URL http://www.seer.cancer.gov/seerstat.

Tufte E (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire.

Tufte E (1990). *Envisioning Information*. Graphics Press, Cheshire.

Tufte E (2006). *Beautiful Evidence*. Graphics Press, Cheshire.

Ventura A (2012). "Micromaps." URL http://support.sas.com/demosdownloads/downarea_t4.jsp?productID=111232&jmpflag=N.

Wainer H (1993). "Tabular Presentation." *Chance*, **6**, 52–56.

Wickham H (2009). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. URL http://had.co.nz/ggplot2/book.

Zhang C (2012). *Interfaces and Visual Analytics for Visualizing Spatio-Temporal Data with Micromaps*. Ph.D. thesis, George Mason University.

Affiliation:

Linda Williams Pickle
StatNet Consulting LLC
20203 Goshen Rd., No. 189
Gaithersburg, MD 20879, United States of America
E-mail: Linda@statnetconsulting.com
URL: http://www.statnetconsulting.com/micromaps.html,
http://mason.gmu.edu/~dcarr/Micromaps/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 63, Issue 3 *Submitted:* 2013-06-01
January 2015 *Accepted:* 2014-10-06