



# Conditioned Choropleth Maps and Hypothesis Generation

Daniel B. Carr,\* Denis White,\*\* Alan M. MacEachren\*\*\*

\*Center for Computational Statistics, George Mason University

\*\*US Environmental Protection Agency, Corvallis, Oregon

\*\*\*GeoVISTA Center, Department of Geography, Pennsylvania State University

The article describes a recently developed template for multivariate data analysis called conditioned choropleth maps (CCmaps). This template is a two-way layout of maps designed to facilitate comparisons. The template can show the association between a dependent variable, as represented in a classed choropleth map, and two potential explanatory variables. The data-analytic objective is to promote better-directed hypothesis generation about the variation of a dependent variable. The CCmap approach does this by partitioning the data into subsets to control the variation in the dependent variable that is associated with two conditioning variables. The interactive implementation of CCmaps introduced here provides dynamically updated map panels and statistics that help in comparing the distributions of conditioned subsets. Patterns evident across subsets indicate the association of conditioning variables with the dependent variable. The patterns lead to hypothesis generation about scientific relationships behind the apparent associations. Spatial patterns evident within individual subsets lead to hypothesis generation that is often mediated by the analyst's knowledge about additional variables. Examples show applications of the methods to health-environment interaction and biodiversity analysis are presented. *Key Words:* geovisualization, exploratory spatial data analysis, choropleth maps, hypothesis generation, partitioning sliders, stratified comparison, dynamic map, statistical annotation.

Exploratory spatial data analysis (ESDA) has become an active area of research for cartographers, statisticians, and others (e.g., Carr et al. 1987; Monmonier 1989; MacDougall 1992; Anselin 1994; Cook et al. 1997; Dykes 1997, 1998; MacEachren et al. 1998; Andrienko and Andrienko 1999). Much of the initial work focused on implementation of standard interactive Exploratory Data Analysis (EDA) methods in a spatial context, usually with a map as one of two or more linked views. Recently, attention has shifted emphasis from implementation of ESDA methods toward the role of these methods in problem solving, hypothesis generation, and knowledge construction (MacEachren et al. 1999; Carr, Wallin, and Carr 2000; Carr, Olsen, et al. 2000; Harrower, MacEachren, and Griffin 2000; Andrienko and Andrienko, 2001a).

Controlling for known or suspected variation in a dependent variable due to associated variables is an important approach for generating better hypotheses about spatial patterns. John Tukey (1976) stressed this approach in a paper entitled *Statistical Mapping: What Should Not be Plotted*. His succinct comment was that "less than fully adjusted variables should not be plotted." When much of the displayed variation in a choropleth map is due to known factors, it is difficult for analysts to relate the unexplained components of variation to ad-

ditional factors. Our visual-cognitive system is ill suited to the task of mentally removing known components of variation, envisioning the patterns in residuals, and relating these patterns to other variables.

Today, such guidance from Tukey and others has been partially taken to heart. Epidemiological graphics routinely show sex- and race-specific rates rather than a composite that mixes these population-based sources of variation. Epidemiologists are well aware that the age distributions of populations being studied influence composite mortality rates. To control the variation of mortality rates due to different age distributions in different regions of study, the variable represented is often either age specific or age adjusted to a common reference distribution of ages.

However, there are often important explanatory variables beyond sex, race, and age. Such variables, sometimes called risk factors, are known to account for part of the observed variation in mortality rates. The *Atlas of United States Mortality* (Pickle et al. 1996) lists known risk factors associated with each type of mortality shown in the atlas. For example, cigarette smoking is a well-known risk factor for lung cancer mortality. Few choropleth maps adjust for such risk factors. This omission inhibits constructive hypothesis generation about additional risk factors. While not providing a full adjustment for

smoking rates, separate maps of lung cancer mortality for nonsmokers and for heavy smokers might be very suggestive. Federal agencies have several reasons for not going further in presenting adjusted maps to the public. The purpose of publications is often to convey statistical summaries rather to promote hypothesis generation. Often, detailed data on risk factors is not available or of poor quality. Even if good data were available for risk factors, agencies would have to address substantial scientific, bureaucratic, and communication issues before presenting risk-factor-adjusted estimates to the public. Nonetheless, when thoughtful analysts obtain estimates for known or conjectured risk factors, they would like to control for risk-factor influence on the primary variable of interest and look more deeply for remaining patterns.

Carr, Wallin, and Carr (2000) described the initial CCmaps concept that provides stratified comparison approach to adjust for two explanatory variables. Similar to studying a choropleth map, analysts using CCmaps treat one variable as a dependent variable and seek to generate hypotheses about its variation in a spatial context. However, when using CCmaps, the analyst also seeks to control the variation of the dependent variable that is associated with two conditioning variables before generating hypotheses. The analyst partitions the regions (i.e., enumeration units) of study into strata (subsets) that CCmaps display in juxtaposed panels of choropleth maps. The regions of each subset appear highlighted in the respective panel of a two-way layout of panels. The other regions also appear in the panel, but with a background color. The analyst then generates hypotheses about differences among subsets that are highlighted in different panels and sharper hypotheses about the spatial patterns appearing within different subsets.

Here, we extend the CCmaps template, describe an interactive implementation that includes fast, multiway, partitioning sliders, and discuss its application in two problem domains (health-environment interaction and biodiversity analysis). The extensions include the addition of dynamic QQplots for comparing distributions plus attention to layout, annotation, and interaction details. We strive to use interactivity and thoughtful design to build an easily traversed bridge between naïve mapping and sophisticated statistical analysis. Readers can try the proposed dynamic conditioning, using shareware available from [www.galaxy.gmu.edu/~dcarr/ccmaps](http://www.galaxy.gmu.edu/~dcarr/ccmaps).

While new, the CCmaps template is related to the work of many previous authors. Below, we indicate several of these connections and indicate the key differences.

Many previous authors have used small multiples of panels to show data summaries under varying conditions. For example, Tufte (1983, 1990, 1997) has repeatedly

advocated the use of small multiples of juxtaposed panels. A common map-based application is to show small multiples for the same variable measured at different times. In mortality studies it is common to display multiple maps of a variable describing different populations indexed by sex, race, and age. In cartography, Bertin initially proposed the idea of interactive small multiples in 1967 (Bertin 1983), but it is only recently that technology has caught up with Bertin's proposal (for an example, see MacEachren et al. 2003). Bertin also proposed small multiples that depict different variables of a potentially related set. Pickle et al. (1996) provide instructive examples of this by showing small choropleth maps of seventeen categories of mortality rates (each relative to the corresponding U.S. rate) for different sex and race combinations. These applications typically have multiple dependent variables that are indexed by time, sex, or other factors, and the small multiples are separate complete maps. The CCmaps methods, as presented here, differ from the above scenarios by having a single dependent variable and by having small multiples that highlight subsets of regions.

The Health-VisB prototype of MacEachren et al. (1998) is closely related to CCmaps. They use dynamic, two-way partitioning of a scatterplot showing mortality rate and one risk factor for the regions. They associate the partitioned regions with a two-way color scheme. A manipulable, four-class map depicts the cross between a binary representation of the dependent variable (mortality rates) and a binary representation of a risk factor. Mortality rates above the user-set threshold are shown in blue and those below the threshold in gray. Light and dark shades represent low and high levels of the risk factor, respectively, resulting in a four-category, choropleth map depiction. They note that their system is an interactive implementation of what Monmonier (1992) calls a cross map. Carr, Olsen, and White (1992) provide an environmental example in the same geovisualization special issue. Health-VisB is more general than the current version of CCmaps in that it provides animation over time. CCmaps is more general in the sense of handling two risk factors. The presentation differs between the two approaches in that Health-VisB uses shades of color to distinguish risk categories while CCmaps uses juxtaposed panels.

The closest predecessor to CCmaps is the Cleveland, Grosse, and Shu (1992) display of two-way, conditioned, juxtaposed scatterplots and other statistical panels. The primary differences are that CCmaps displays maps and provides dynamic partitioning. While the idea of conditioned maps (Carr, Wallin, and Carr, 2000; Brunson 2001) is a small step from their previous work, we have devoted substantial attention to developing a thoughtful interactive implementation.

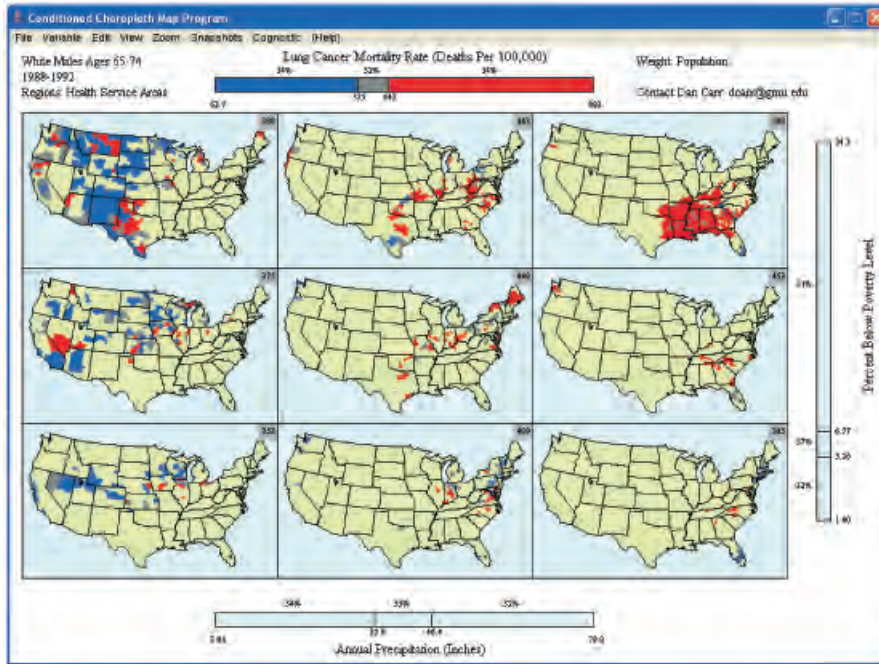


Figure 1. A conditioned choropleth map.

Cleveland, Grosse, and Shu (1992) also introduce a noteworthy generalization to conditioning. They use overlapping intervals called shingles. With overlapping inclusion intervals, cases can appear in more than one panel. This is particularly helpful in reducing edge effects when modeling or smoothing the data within individual panels. To keep things simple, we emphasize strict partitioning of region subsets. Each region appears as a highlighted region in exactly one panel.

The subsequent sections of the article are as follows: First, we provide a detailed description of CCmaps, using an application to health data analysis to illustrate the core components of the method. This section also covers several recent extensions, their implementation, and issues that must be considered for effective application of CCmaps. Next, we use a second case study, an environmental application, to illustrate the potential of the CCmap template. Then, we review design and data considerations for CCmaps. Finally, we close with remarks focusing on extensions to methods and new views, such as cognostics (computer guided diagnostics) and smoothing, respectively. Some of these anticipated changes are now a part of the most recent CCmaps distribution (see the URL above).

## Dynamic CCmaps: Method and Application to Health Data Analysis

CCmaps implements a dynamic map approach that provides analysts with direct, coarse control for the

variation of a dependent variable due to two additional variables. This coarse control derives from using values of two additional variables to partition the regions of study into a two-way layout of panels as shown in figure 1. We call the two variables *conditioning variables*. The task of picking conditioning variables is analogous to picking explanatory, predictor, or independent variables for a regression model. Picking variables is often a tentative first step. If variables don't help to reduce (control) the variation in the dependent variable, we typically drop them or search for alternatives.

The regions of study in figure 1 are health service areas (HSAs). HSAs are counties or aggregates of counties based on where people get their hospital care. These regions formed the geographic base for the mortality atlas cited above and are commonly used in U.S. mortality rate descriptions produced by the National Center for Health Statistics.

The dynamic version of CCmaps presented here introduces partitioning sliders that allow an analyst to partition regions dynamically into three classes—low, middle, and high—based on the values of an ordered numeric variable. The partitioning slider at the bottom partitions the HSAs in figure 1 into subsets with low, medium, and high values based on precipitation. The slider at the right partitions HSAs into three subsets or classes based on the percent of households below the poverty level. The cross-product of these two univariate partitions leads to the definition of nine subsets.

CCmaps highlights these nine subsets in a  $3 \times 3$  grid of panels. Regions with low, medium, and high values of annual precipitation appear in the left, middle, and right columns, respectively. Regions with low, moderate, and high values of percent poverty appear in the bottom, middle, and top rows, respectively.

The slider at the top also partitions the regions into three classes based on values of the dependent variable, lung cancer mortality. Regions with low, medium, and high values have colors of blue, gray, and red, respectively. To provide the context of a full map in each panel, CCmaps displays all of the regions. When the region does not meet the selection criteria for a panel, CCmaps changes the region color to the background color, a desaturated yellow. When we refer to regions in a panel, below, it is to be understood these are regions that have met the selection criteria and appear as blue, gray, or red.

The partitioning of regions can produce subsets that have much less variation in the dependent variable than the collection of all regions. In figure 1 the regions in the top right panel have high values for both precipitation and poverty. Most of the regions are red, indicating high values for lung cancer mortality. The conditioning has removed most of the regions with low and moderate mortality rates from the panel and produced a subset that is more homogeneous in terms of lung cancer mortality.

The CCmaps' partitioning sliders have two inner, mouse-adjustable boundaries shown as line segments. These two inner boundaries control three-way classification. The slider outer boundaries indicate the slider scale limits. CCmaps sets the scale limits to include all regions.

We proposed CCmaps as a way of drawing practitioners, as well as students, toward a deeper understanding of multivariate spatial data. Counting the spatial indices as two dimensions, the use of CCmaps allow analysts to look for relationships in five-dimensional data using nothing more complicated than three sliders connected to a choropleth map. Limiting what is depicted in each panel and separating the within-panel comparison from a modest number of between-panel comparisons has the potential to reduce the perceptual and cognitive tasks of 5D analysis to manageable ones.

CCmaps provides an easy but coarse way of controlling variation of a dependent variable and studying the results in the context of a map. Our goal in providing CCmaps as a dynamic application is to encourage a broad audience to take simple next steps toward the generation of better hypotheses about the observed variation of a variable. Toward this goal, we include some simple statistics, such as weighted means, to help in

comparing the distributions of the subsets defined by partitioning sliders.

More refined but less broadly accessible statistical methods are available for studying variation of dependent variables. In particular, known explanatory variables can be included in sophisticated models. CCmaps can then display model values or model residuals to assess possible associations with two additional variables. Choices for conditioning variables can also include estimates of uncertainty, influence, and other model diagnostic statistics. The patterns observed can lead to hypotheses about model inadequacies.

When examining spatially indexed estimates, model values, or residuals, the use of maps is often helpful for hypothesis generation because people organize portions of their background knowledge spatially. Showing values in a spatial context can reveal spatial patterns. The potential for insight increases when analysts can associate these patterns with other variables.

When looking at CCmaps, analysts are often interested in multiple questions. Four general questions are:

1. Are the distributions of values in the panels different from each other?
2. If so, are the conditioning variables "causal variables," or are they partial surrogates for more scientifically plausible causal variables?
3. Is the variation of values in an individual panel different than noise?
4. If so, what variables or facts might account for part of the variation?

Questions 1 and 3 are screening questions. These are important in that analysts do not want to waste time generating hypotheses about variation that is essentially noise. The current version of CCmaps avoids the complications of testing rigorous hypotheses about these two questions. However, CCmaps provides descriptive statistics that are helpful in regard to question 1. As a start, this section of the article introduces the weighted average of region values as annotation for each map panel. Later, we go further to describe QQ plots that provide a more thorough visual comparison of distributions.

Questions 2 and 4 are hypothesis generation questions. The hypotheses generated depend substantially upon the knowledge that analysts bring to the problem. The strong association of precipitation with lung cancer mortality in figure 1 may surprise many people at first glance. There is no known scientific mechanism for precipitation directly causing lung cancer. Thus, precipitation is likely to be serving as a surrogate for, or be confounded with, causal variables.

Below, we address the notions of surrogate and confounded variables in greater detail. First, however, we describe statistical annotation features in the current implementation.

### Statistical Annotation for Map Panels

Annotation improves our ability to interpret maps (see Bertin 1983; MacEachren et al. 1998). To encourage statistical thinking, we supplement the map panels and sliders with weighted descriptive statistics. At the top right of each panel, we show the weighted average of values for the highlighted regions. We also calculate percents for the weights that are in categories defined by each slider. The percents appear at the center of the intervals indicating the category boundaries.

The preferred choice of weights depends on the application. We are not usually content with the common GIS default that gives regions equal weight. Figure 1 focuses on the mortality rates of people living in the continental United States. In this context our preference is to give people equal weight. Since figure 1 covers a five-year period, we use the number of person-years at risk (basically, the number of white males within the given age range living in a region times five years) as weights when producing a panel summary. The weighted average rate for regions in a panel then has a simple interpretation. It is the total deaths over five years summed over the regions in the panel divided by the total person-years at risk summed over the regions in the panel.

When visually comparing map panels in terms of lower and higher mortality rates, we often want to factor in the number of people involved, but our visual impression can be dominated by our visual response to the area of colored regions. Our perceptual/cognitive systems automatically give more weight to regions with larger areas whether or not area is appropriate as a weighting factor. In terms of a single value for comparing panels, the logical choice is a person-years-at-risk weighted average for the highlighted regions, not a region-area weighted average. Our visual impression is also limited by having access to just the three class-encoded mortality rates and not the original values. Thus the choropleth maps only provide rough impressions of mortality rate distributions, and annotation can be important for a more appropriate interpretation.

Still, rough impressions can often steer us in the right direction. In looking at the top left and top right panels in figure 1, we see at a glance a considerable amount of blue in the left panel and a preponderance of red in the right panel. We correctly infer that the high mortality

rates are concentrated in the right panel and that the weighted average rate is higher in this panel. We can then refine our visual impression by attending to statistical summaries. The values in the corresponding tabs are 380 and 508 deaths per 100,000. The difference between the rates of 128 deaths per 100,000 is highly meaningful in an epidemiological context. Further steps might be to test the statistical hypothesis of the same mortality rate for all the panels and to test for differences among various subsets of panels. Addressing statistical tests and multiple comparison issues raised by both multiple panels and dynamic partitioning is beyond scope of this article. In brief, CCmaps provides a fast but rough, color-based, visual comparison and a slower, text-based, but more quantitative comparison of panel average rates.

The sliders in CCmaps also use weights in computing the percents associated with each partitioning interval. The top slider shows that 34, 32, and 34 percent of the people in the study are associated with the low, medium, and high partitioning intervals, respectively. In other words, 34 percent of the person-years at risk are associated with regions colored blue.

In our interactive exploration we often like to start with roughly one-third or 33 percent for each category. When the two partitioning variables are independent, this tends to balance the person-years risk in the nine panels. The uncertainty of panel rate estimates increases as the number of person-years at risk decreases. The balancing tends to avoid instances of relatively high uncertainty. It is easier to think about panel differences when we are expecting  $1/3 * 1/3 = 1/9$  of the person-years in each panel. Of course the conditioning variables may not be independent, so the panel person-years at risk can be quite unbalanced even though the sliders set the margin percents at about 33 percent. (CCmaps presents individual panel percents in the QQ plots view as discussed below.) In practice it may be impossible to obtain a targeted percent for a slider category since inclusion of just one region can change the population percents by values larger than one. We are only suggesting equal margin percents as a starting place for exploration. Later, we describe a fast way to find slider settings so that the weighted means for the panels provide a better fit to the regions values in the panels.

### Variable Selection and Hypothesis Generation for Figure 1

This section addresses variable selection and hypothesis generation in the context of the single example

shown in figure 1. The variables chosen for figure 1, lung cancer mortality rates, precipitation amounts, and poverty percents, seem an unlikely combination. What has precipitation got to do with lung cancer? We purposely chose this example to discuss issues related to the choice of variables.

In terms of variable selection, a major difficulty in using CC-maps for a desired study is likely to be lack of data, lack of access to data, or lack of quality data. When the variables that would be the most meaningful in a rigorous scientific framework are not available, researchers often choose to use surrogate variables in order to make some progress. The less than perfect associations between the surrogate variables and the desired variables complicates interpretation, but surrogates may serve adequately in the context of hypothesis generation.

### The Geographic Regions

The selection of spatial resolution can influence the availability of data. The smaller the regions, the more likely it is that good estimates are not available or that available estimates produced by federal agencies are inaccessible due to confidentiality considerations. At the same time, aggregation to larger regions can hide the spatial variability that is of crucial interest.

As indicated previously, the regions shown in all the map panels of figure 1 are Health Service Areas (HSAs). The map projection is the National Center for Health Statistics standard, an Albers conical equal area projection. It is sometimes helpful to show the region outlines when closely studying an enlarged view; however, to avoid visual clutter our default menu option hides the HSA outlines. In CCmaps, clicking on a panel enlarges it, and clicking again returns to the  $3 \times 3$  layout.

### The Variables

The gathering of variables shown in figure 1 arose in the context of an exploratory study of chronic obstructive pulmonary disease (COPD) mortality rates. While lung cancer mortality is the dependent variable in figure 1, all three variables in the figure were initially treated as risk factors in the COPD study. The following description motivates the collection variables for the COPD study and indicates how the strong association between lung cancer mortality and precipitation was discovered.

COPD is a broad category that includes many specific causes of death such as emphysema, asthma, and bronchitis. The risk factors listed in the mortality atlas include cigarette smoking, coal dust exposure, and low socioeconomic status as related to poor housing. In the effort to

obtain risk factor data quickly, we decided to try readily available surrogates. We chose lung cancer mortality as a surrogate for cigarette-smoking rates. This surrogate risk factor for the COPD study became the dependent variable in figure 1. We selected percent of households below the poverty level as a surrogate for low socioeconomic status, poor housing, and limited access to medical services. This remained a risk factor for figure 1.

An examination of the COPD maps in the mortality atlas motivated our selection of the third risk factor for consideration, precipitation. Higher COPD mortality rates seemed noticeable in the dusty western HSAs. Since coal mining was already identified as a risk factor, it seemed reasonable to conjecture that inhalation of particulates other than those associated with coal mining might be a risk factor. However, particulate monitoring in the United States is mostly limited to Metropolitan Statistical Areas. We were not aware of particulate data available on national scale. We conjectured that particulate exposure might be inversely associated with precipitation. Precipitation data are readily available (National Climatic Data Center 2004). Since the data were for meteorology stations and not HSAs, we adapted them to HSAs. Our calculation procedure associated each of the thousands of weather stations with the closest HSA centroid using great circle distance. We then computed a precipitation estimate for each HSA as the median of the associated annual precipitation amounts over the period 1988–1992. A scatterplot of COPD mortality rates versus precipitation is very noisy. A weak pattern shows though the noise: HSAs with moderate precipitation tended to have lower COPD mortality rates. A scatterplot matrix of COPD mortality rates and the three risk factors revealed the strong association of precipitation with lung cancer mortality rates. This discovery led to the CCmaps shown here as figure 1.

### Tabular Evaluation

Statistical tests of hypotheses can be applied to the population-weighted average rates appearing in the top right of the figure 1 panels and in the body of Table 1a. Since the underlying data consist of counts—deaths and number of person-years at risk—it is possible to reexpress the  $3 \times 3$  weighted averages as a  $2 \times 3 \times 3$  table and proceed to test hypotheses about the table using the Mantel-Haenszel statistic and other count-based test statistics. In general, the variable of interest may take different forms, so analysis of variance, log linear models, and other methods may provide frameworks for hypothesis tests. The simple application of such tests ignores spatial variation, does not include our additional

**Table 1a.** Weighted Means

Poverty	Precipitation			Ave
	Low	Middle	High	
High	380	443	508	444
Middle	375	447	453	425
Low	353	417	383	384
<b>Ave</b>	<b>369</b>	<b>436</b>	<b>448</b>	<b>418</b>

knowledge, and does not account for the multiple comparisons generated by adjusting sliders to produce different tables. Adjusting for multiple comparisons in significance tests can help to guide us away from generating hypothesis about what later turns out to be noise. However, marginally relevant significance tests are only a rough guide in a hypothesis generation context. When generating hypotheses, patterns of practical significance are sufficient to warrant further consideration.

Our preference for quick inspection is to characterize the two-way table of weighted averages in terms of main effects and interactions as in analysis of variance. Table 1a shows the rates in terms of deaths per 100,000, and Table 1b reexpresses them as a grand average, row and column effects, and interactions. The grand average for the table is 418 deaths per 100,000. The precipitation main effects are roughly  $-48$ ,  $18$ , and  $30$  deaths per 10,000. In terms of average values, the mortality rates for regions associated with the right column are about 78 deaths per 100,000 higher than the region rates associated with the left column. Similarly, the percent of households below the poverty level main effects are roughly  $-33$ ,  $7$ , and  $26$  deaths per 100,000. On the average, the rates for regions associated with the top row are about 59 deaths per 100,000 higher than region rates associated with the bottom row. The magnitudes of some interactions in the body of Table 1b rival the main effects. For example, the cell value for high precipitation and high poverty is 34. The presence of extreme positive or negative interactions warns that a description in terms of simple main effects fails to tell the full story. The two-way table basically confirms what is visible by a glance

**Table 1b.** Main Effects and Interactions

Poverty	Precipitation			Effect
	Low	Middle	High	
High	$-15$	$-19$	$34$	$26$
Middle	$-2$	$4$	$-2$	$7$
Low	$17$	$15$	$-32$	$-33$
<b>Effect</b>	<b><math>-48</math></b>	<b><math>18</math></b>	<b><math>30</math></b>	<b><math>418</math></b>

at the maps. There is a lot of blue in the left column of panels (locations with low precipitation) and a lot of red in the top right panel (locations with high precipitation and high percent below the poverty level). The structure appearing in the CCmaps, the tables, and the QQplots (not shown) seems unlikely to be generated by noise.

As for calculations in Table 1b, the row and column effects are simply the row and column averages in Table 1a minus the grand average. The interactions in the body of Table 1b are the Table 1a cell values minus the grand average and the respective row and column effects. For simplicity we have skipped the use of panel weights and rounded to integers. (Some analysts might prefer a median polish, described in Emerson and Hoglin 1983, so the high value in the top right contributes less to its row and column margins.) CCmaps includes a menu item to show tabular views that include a table of means and a table of main effects and interactions.

### Data Quality, Hypotheses Generation and Confirmation

Exploratory epidemiology studies like those described here are often called ecological studies. In ecological studies the variables characterize populations associated with a collection of geographical regions. While other kinds of population studies, such as case-control studies and cohort studies, can also have their problems, the statistical foundations of ecological studies are often suspect (e.g., see Richardson and Monfort 2000 or Lawson 2001). The limited quality and indirect relevance of available data is often incompatible with a coherent model or cause and effect conclusions concerning well-defined populations. Difficulties are sometimes referred to as “the ecological bias or fallacy.” Some may even argue that it is a waste of time to investigate such data. Our position is that it is good to have thoughtful people explore health and environmental data, even when the data are not ideal. Sometimes the suggestive patterns turn out to be real.

Those engaging in hypothesis generation need to be aware that:

1. Randomly generated patterns often appear to have structure;
2. Humans have remarkable ability to generate explanations;
3. Many datasets really do have deficiencies in quality or appropriateness.

In terms of data problems we know at the start that:

1. Mortality rates have problems with inconsistent reporting and errors;

2. Genetic susceptibility is often an important but unmeasured risk factor;
3. Exposure estimates summarized at a region level are at best a primitive guide to the temporal exposure patterns of individuals and subpopulations within the region;
4. The available exposure estimates may be poor surrogates for the most relevant time periods in a model relating exposure to observable response;
5. Due to migration, the population associated with a region is not a static entity. Different kinds of region statistics characterize different populations that have only partial overlap;
6. Available estimates can be biased or have large uncertainty;
7. There are many uncontrolled variables and many potential confounded variables.

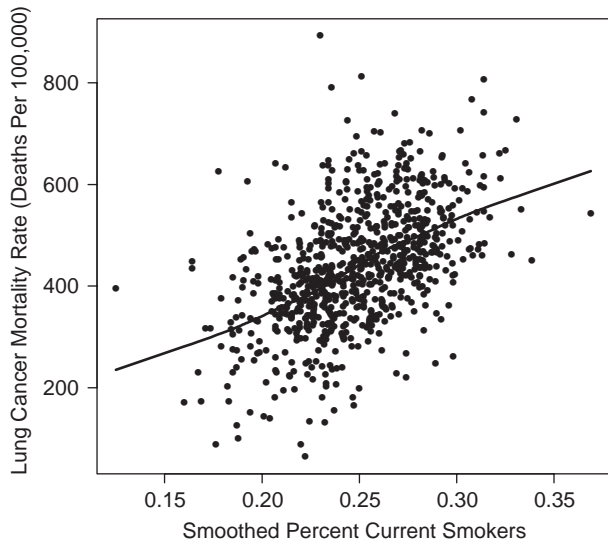
When patterns show through all the noise, we encourage people to apply their domain knowledge and understanding of the spatial context to conjecture about credible mechanisms behind the observed patterns. The next step may be to talk with experts or to seek new data that help in evaluating the conjecture explanation. Firm confirmation is typically a slow process involving consensus of the many scientists using other studies, such as case control and laboratory studies.

In our lung cancer mortality example, precipitation amounts are highly associated with lung cancer mortality. This surprising pattern begs explanation, providing a prototypical hypothesis-generation situation. The top right panel shows high mortality rates in the Southeast. Many of the HSAs in the Southeast have high tobacco production. The heavy influence of the tobacco industry over decades has undoubtedly influenced the local culture toward increased cigarette consumption. This cultural influence surely extends well beyond just the HSAs engaged in tobacco production. A reasonable conjecture is that precipitation amounts are partially confounded with cigarette smoking over broad areas of the Southeast. It is also possible that there might be other behavioral connections to cigarette smoking beside the tobacco economy of the Southeast. Perhaps cigarette smoking is more irritating in dry dusty environments, and people tend to smoke less there. We have little doubt that cigarette smoking is the major factor in the variation in lung cancer mortality rates.

Poverty also helps to mark out the Southeast. In addition, poverty could be associated with limited response to antismoking education and hence associated with greater cigarette smoke exposure. It is, thus, plausible that both precipitation and poverty are partially confounded with cigarette-smoking rates.

To get support for our cigarette-smoking hypothesis, a natural step would seem to be getting data on cigarette-smoking rates. Amazing as it may seem, good estimates of cigarette smoking rates at finer resolution than the state level are not available. Available national survey data that include questions about tobacco usage include: NHIS (National Health Interview Survey: adult core and cancer control surveys), NHANES (National Health and Nutrition Examination Survey: family and sample person surveys), CPS (Current Population Survey: core tobacco and cessation surveys), BRFSS (Behavioral Risk Factors Surveillance System), and NHSDA (National Household Survey on Drug Abuse). Questions such as “Have you ever smoked?” provide primitive information. We would rather have answers to questions about the years of smoking and average packs per day. While these questions may be hard to answer accurately, they would seem to provide a better foundation for determining the dose distributions for the different regions. BRFSS does have some supplemental questions related to smoking rates, but these are not asked in all states for all years. None of the surveys were designed to produce estimates for HSAs or counties. The desired data are not at hand.

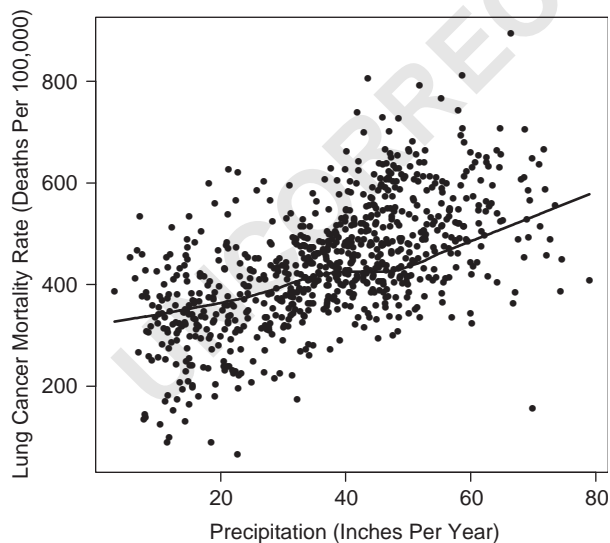
To at least look at some data on smoking we turned to Pickle and Yuchen (2002), who compiled risk factors at the county level based on BRFSS survey data. While little can be done to address the bias that may be present due to unrepresentative sampling, they use spatial smoothing to strengthen estimates that may be poor due to small sample sizes. We used their smoothed values for the percent of males answering “yes” to the “ever smoked” and “current smoker” questions. Figure 2a, based on HSA estimates, is a scatterplot with a smoothing-based estimate of a functional relationship. It shows the limited association of the percent males (who were current smokers time of the survey), smoothed values, and the lung cancer mortality rates. The correlation is only 0.51. Likely, the answer to the current-smoker question is a poor surrogate measure of the relevant exposure. We also know the mortality rate population and the surveyed population differ in terms of age. There is also a time-frame mismatch. The previous twenty years of cigarette smoking is relevant to lung cancer mortality. A recent study does not necessarily reflect the patterns of a previous decade or two. We are not surprised by the weak association in Figure 2a and continue to think that cigarette smoking is the major factor behind the spatial pattern. With less than national county level assessment in mind, we have initiated efforts to obtain smoking-rate estimates from the CPS and BRFSS surveys. The CPS rates will be limited to the



**Figure 2a.** A scatterplot using percent current smokers and augmented by a population-weighted smooth. The unweighted correlation is 0.51.

counties in the metropolitan statistical areas. The coverage from BRFSS will depend on the states that chose to incorporate the supplemental smoke rating questions.

Figure 2b shows the scatterplot with precipitation as the risk factor. While there is a lot of noise, the smooth curve is strongly suggestive. When in hypothesis-generation mode, it is easy to stop at the first plausible explanation. While at risk of being deceived by patterns unadjusted for cigarette-smoking rates, we proceed with hypothesis generation. Perhaps the somewhat stronger association (correlation 0.60) of lung cancer mortality



**Figure 2b.** A scatterplot using precipitation and augmented by a population-weighted smooth. The unweighted correlation is 0.60.

rates with precipitation is not totally through the conjectured connection to cigarette-smoking rates. Perhaps there is something more going on than just an association between locations with high rates of cigarette smoking and high precipitation. Before speculating about other factors related to precipitation that may also be related to lung cancer mortality, it is appropriate to review the weaknesses of the data in view of the above list. Some of our thoughts are as follows:

In terms of figure 1, the reporting of lung cancer as a cause of death was being handled pretty uniformly by 1988. Many years earlier, the reported cause of death could have been pneumonia, a metastatic cancer, or other problems related to the lung cancer. Retirement is a transition time when many people consider moving. Thus, many men in the age group 65–74 may have changed location. Many of the people dying could have come from regions with quite different descriptors for precipitation and poverty. However, we conjecture that migration is not likely to be a major factor. Wealthy people may be able to seek their preferred climate and living communities. Reasonable conjectures are that migration of the wealthy to live in the poor HSAs in the southeastern United States is limited and that a limited number of poor in the southeastern United States move to HSAs with radically different values for the three variables. While further information would be helpful, this situation may not be as troublesome as situations in which there is less regional similarity.

If precipitation rates change substantially over time and are somehow associated with some lung-cancer-causing mechanism, perhaps the time period for the precipitation data is poorly picked. Since twenty years of cigarette smoking is related to lung mortality, perhaps earlier precipitation would be more relevant. How could precipitation be related to lung-cancer mortality rates in a way other than through its relationship to cigarette smoking? Other recognized causes of lung cancer include asbestos exposure and radon exposure. Moisture can decrease the air transport of asbestos fibers, but that would seem to be a very local phenomenon. It seems likely that this specific mechanism would tend to decrease mortality rates if it were a noticeable factor. In some places, rainfall could encourage people to stay inside where there are higher levels of radon, cigarette smoke, and, possibly, asbestos fibers or other indoor hazards. Of particular interest to us is the connection of precipitation to humidity and humidity to other factors. An early thought was to look for studies related to molds. One researcher (personal communication) suggested that this line of inquiry would not be fruitful. Molds have been associated with asthma but not with

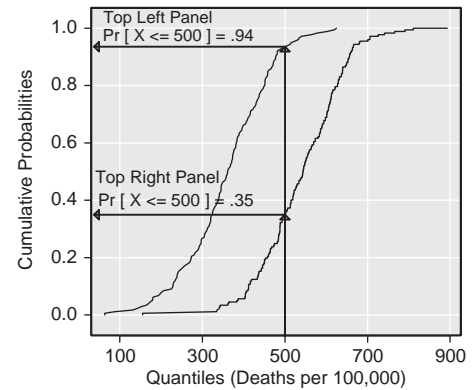
carcinogenicity. Another researcher (personal communication) suggested that humidity in the lungs might increase the chemical activity of hazardous air pollutants on the lung surfaces. There are many possible indirect connections between precipitation and lung cancer mortality. The lack of good data on cigarette-smoking rates prevents conditioning to control this known source of variation and represents a major barrier to moving from armchair speculation to more serious hypothesis generation.

## QQplots

One objective of partitioning regions into juxtaposed panels is to foster distributional comparisons for the subsets created. The weighted average cited above is often useful as a summary statistic describing a distribution. However, two radically different distributions can have the same means, so focusing only on the averages (the estimates in the CCmap tabs) can miss the real story. Any single summary statistic may do a poor job of characterizing a distribution of values and may provide a poor basis for making comparisons. Box plots provide a richer distributional caricature by showing the median, quartiles, and extrema. (There are many variations on box plots. Some, for example, show adjacent values and outliers rather than extrema.) The cumulative distribution tells the full univariate story. Thus, detailed comparisons are often based on cumulative distributions (Cleveland 1993; Andrienko and Andrienko 2001b) Quantile-quantile plots (QQplots), invented by Wilk and Gnanadesikan (1968) and described in more detail below, allow analysts to compare quantiles of two cumulative distributions for matched cumulative probabilities. Thus CCmaps provides dynamic QQplots to encourage incisive distributional comparisons.

QQplots are not widely used, so some explanation is appropriate. A good starting point begins with a description of cumulative distribution plots.

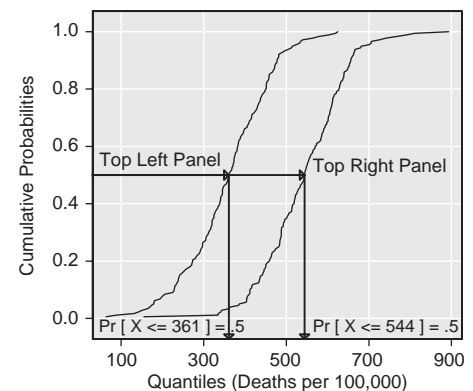
The notions of cumulative probabilities,  $p$ , and quantiles,  $q$ , are fundamental to understanding the construction of cumulative distribution plots. Let  $p = Pr(X \leq q)$ . The empirical cumulative distribution provides one basis for estimating pairs  $(p, q)$ . Given any value  $q$  (a quantile), let the estimated value for the cumulative probability  $p$  be the number of cases with values less than or equal to  $q$  divided by the total number of cases. The empirical cumulative distribution is a step function with values for  $p$  running from 0 to 1 and a domain for  $q$  consisting of all real numbers. For typical cumulative distribution plots, the domain covers just the range of observed values and the extension to plus and minus



**Figure 3a.** Two CDF curves: obtaining probabilities given a quantile.

infinity is implicit. In a plot we can overlay two empirical cumulative distributions to make comparisons. We can pick a quantile and compare the two cumulative probabilities. Figure 3a shows an example comparing the distributions for the regions in the top left panel and the top right panel of figure 1.

There are two drawbacks to assessing such a graphic directly. First, it requires assessing the vertical distance between two curves, a task not readily performed by our eye-brain system (Cleveland 1985). Second, interpretation is a bit tricky because the curve with the higher cumulative probabilities has more small values. Suppose we compare cumulative percents for two distributions at a mortality rate of five hundred per one hundred thousand. As shown in Figure 3a, the cumulative probabilities are 0.94 and 0.35 (94 percent and 35 percent) for the top left panel of the CCmaps (low precipitation) and top right (high precipitation) distributions, respectively. Therefore, 94 percent of the HSA mortality rates in the top left panel are less than or equal to five hundred per one hundred thousand, whereas only 35 percent of the top right panel rates are that low. This *larger* probability for



**Figure 3b.** Two CDF curves: obtaining quantiles given a probability.

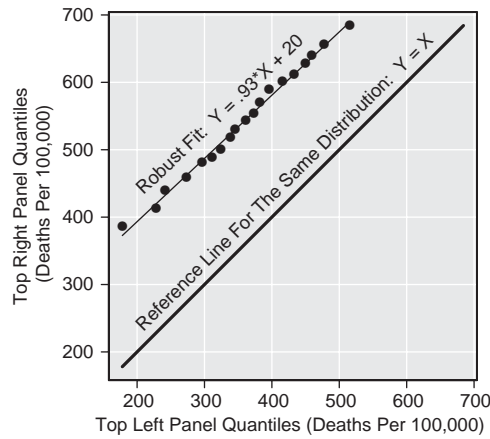


Figure 3c. A QQ plot with a robust straight-line fit.

low precipitation regions means the lung cancer mortality values in these regions tend to be *smaller*. Making a mental switch from larger to smaller can be confusing.

For this and other reasons, many analysts prefer to compare the quantiles for the same cumulative probabilities. Figure 3b illustrates the basic idea and provides the foundation for constructing QQplots. In accordance with Cleveland's (1993) guidance, Figure 3b modifies the construction of the cumulative distribution curves by interpolating percents between adjacent sorted observations. Typically, there are no ties in observations for estimated mortality rates, and the piecewise linear curve is a monotone-increasing curve that has an inverse. Thus, we can select a probability, such as 0.5, and obtain the corresponding unique quantile from each of the two distributions, as indicated in the Figure 3b. The result is a pair of quantiles, (361, 544).

Figure 3c is a scatterplot of quantile pairs determined using a set of probabilities: 0.05, 0.1, . . . , 0.95. The first element of a pair comes from the top left panel distribution and is plotted on the x-axis, and the second element comes from top right panel distribution and is plotted on the y-axis. The thin line in the Figure 3c shows a robust straight line fit to the points. If, as in this case, the line fits the points reasonably well, the distributions can be meaningfully compared based on their means and standard deviations. When points are close to a straight line, the higher moments of the distributions are similar. The slope of the robust line then reflects the ratio of standard deviations. In the example, the ratio is 0.93. If the ratio (and slope) is 1, the difference in means can be determined directly from the fitted line's intercept, and the difference in means tells the whole story of distribution differences. Figure 3c indicates that the means provide a reasonably good basis for comparing these two subsets.

There are some differences between the current CCmaps implementation than that suggested by the figure 3 examples. First, the cumulative distributions in CCmaps now incorporate population weights. Second, the distribution in each panel is compared to the composite distribution of all other panels. Third, the number of *qq* pairs in a panel does not always reflect the smaller of two sample sizes. When the smaller sample has more than one hundred regions, CCmaps approximates one hundred quantiles from both distributions using a common set of cumulative probabilities. This is adequate detail for most purposes. Finally, CCmaps augments the QQplots with a single bivariate point showing the paired minima of the set and its complement. The interpolation process does not naturally pair the minima since they are associated with different cumulative probabilities. The maxima are naturally paired since they are both associated with cumulative probabilities of 1. Seeing both the minima and maximum helps in locating the regions with both kinds of extreme values among the panels.

We augment the QQplots panels with statistics that include the weighted average, the sum of weights for the individual panel, and the percent of total weights. For example, the top left panel in figure 1 represents 1.3 million person-years of experience, and that is 10 percent of the total person-years.

In summary, QQ plots have the merit of revealing distributional shape discrepancies as departures from a straight line. When there are major shape discrepancies, simple descriptive statistics such as means and standard deviations provide a questionable basis for comparing distributions. When the shape discrepancies are modest, QQ plots can be interpreted in terms of comparing standard deviations and means.

## An Environmental Example

The CCmap method is not limited to use with traditional choropleth maps based on political enumeration units such as the HSAs used above. Here we illustrate the potential applicability of the method to an environmental science application using a regular grid for integrating data.

Environmental data often have complicated, multivariate, nonlinear relationships. We need good tools to help us to see patterns and think about possible explanations. Exploratory fitting techniques such as regression trees (O'Connor et al. 1996) help by weeding out variables not related to the dependent variable and by bringing out structure that is often hidden by random variation. Both regression trees (White and Sifneos 2002) and CCmaps address nonlinearity through parti-

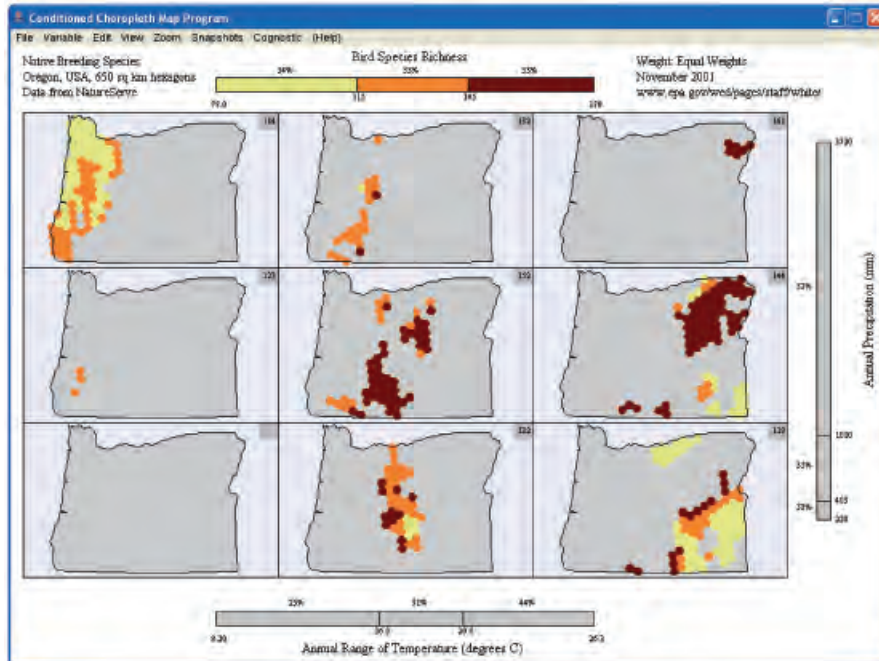


Figure 4. A CCmap of bird species richness conditioned by annual range of temperature and annual precipitation.

tioning the data into domains where simpler models can provide a more adequate description and it is easier to generate reasonable hypotheses about the patterns observed.

As an example, biogeographers have long sought explanations for spatial patterns of species richness (Brown and Lomolino 1998). We illustrate two cases of these patterns with CCmaps in figures 4 and 5 using bird

species richness (numbers of species) as the dependent variable and, for the first map, annual range of temperature and annual precipitation as the conditioning variables and, for the second map, mean annual temperature and range of elevation.

These data were prepared as part of a collaborative biodiversity research project that has been studying how biodiversity responses such as species richness are

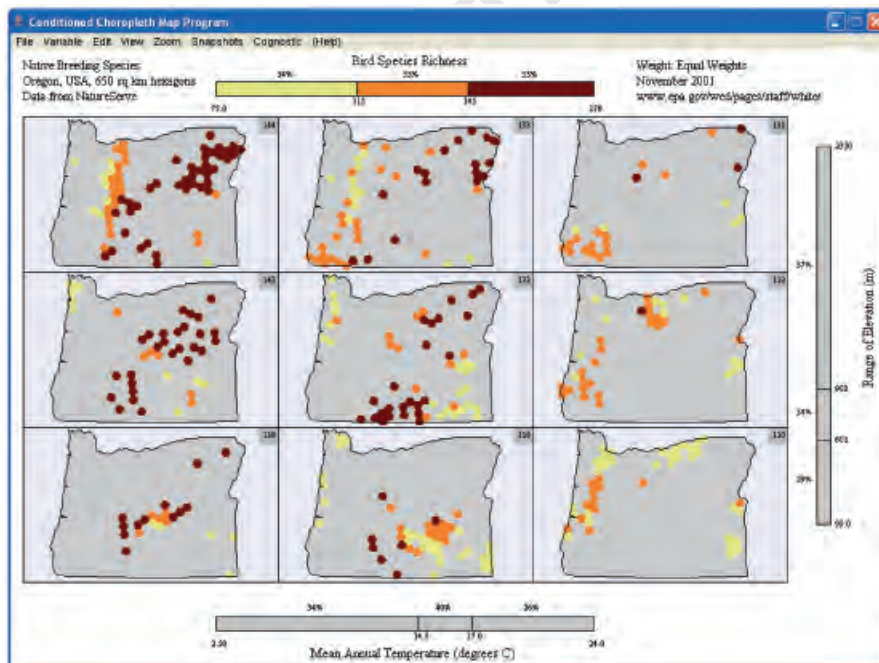


Figure 5. A CCmap of bird species richness conditioned by mean annual temperature and range of elevation.

associated with environmental factors, how to best prioritize areas for conservation, and how biodiversity responses may be impacted by future landscape change (White et al. 1999). For this study, Oregon is divided into a regular grid of hexagon cells approximately 650 km<sup>2</sup> in size (White, Kimerling, and Overton 1992). The size of the grid cells reflects a compromise between the desire for spatial detail, on the one hand, and the constraints of reasonable spatial representation of species life histories, of data collection, of confidentiality, and of computational feasibility, on the other hand.

The Nature Conservancy prepared bird species distributions from standard range information in published literature, from expert sources, and from specific location data on species of conservation concern (Master 1996). Topographic elevation, temperature, and precipitation data were compiled to a rectangular grid at a resolution of 1 kilometer for the conterminous United States. The elevation data were derived from a 15 arc-second digital elevation model. Temperature data were modeled and compiled using the method of Marks (1990). Precipitation data were compiled from data prepared by Daly, Neilson, and Phillips (1994). The 1 km data were aggregated and summarized by 650 km<sup>2</sup> grid cell (White et al. 1999).

Over the earth as a whole, those areas that are warmer, wetter, less seasonal in temperature, and more varied in topography are generally associated with higher species richness (Caldecott et al. 1996). To examine these relationships in Oregon, we first look at scatterplots of the conditioning variables with the dependent variable (figure 6). No strong relationships are suggested by these plots, though annual range of temperature and annual precipitation have a somewhat more structured relationship than the other two variables. The envelope type of relationships of both annual range of temperature and annual precipitation with bird-species richness are not unusual in environmental data and suggest mechanisms that act more as thresholds than as gradients.

The spatial patterns in the conditioning variables are quite marked, however (figure 7). Annual range of temperature has a distinct, nearly linear, gradient from west to east across Oregon. Annual precipitation has almost the reverse, except that higher precipitation in eastern Oregon is found in the mountainous areas rather than in the most eastern areas. Mean annual temperatures are highest in the valleys of western Oregon, the basins of eastern Oregon, and in southwest Oregon. Range of elevation is highest in the Cascade Mountains where, indeed, the highest maximum elevations are also

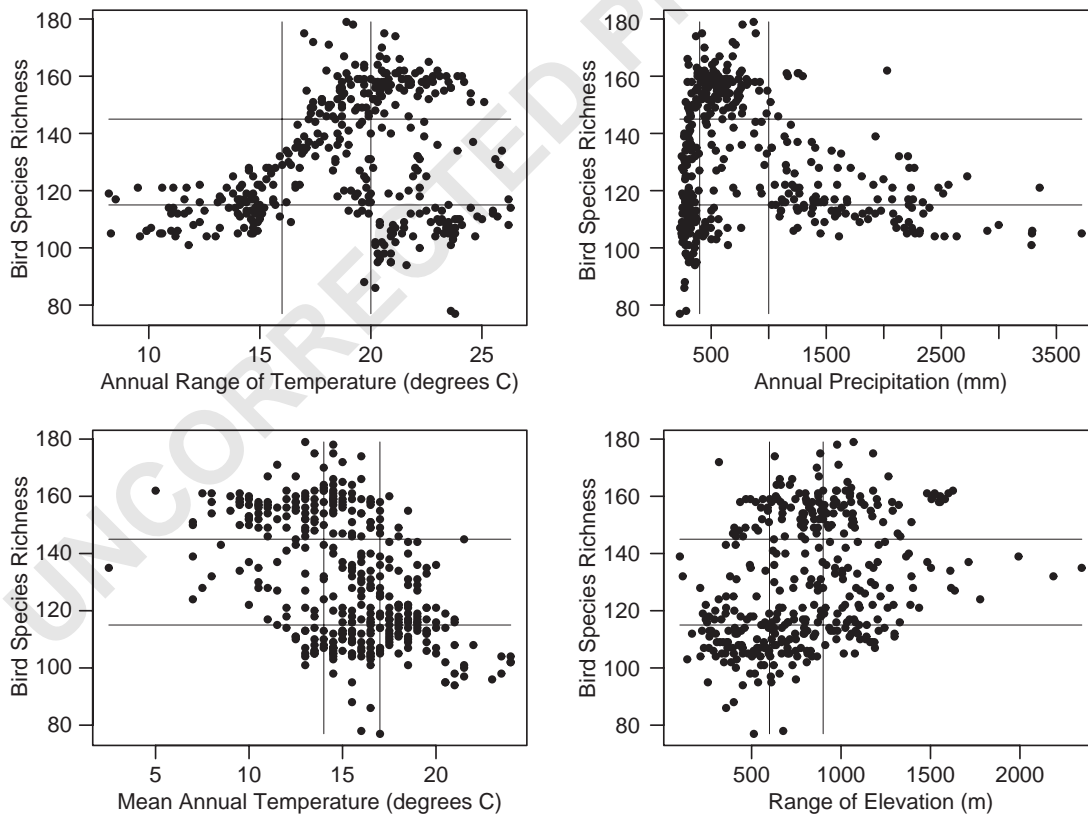


Figure 6. Partitioned scatterplots for a dependent variable and conditioning variables.

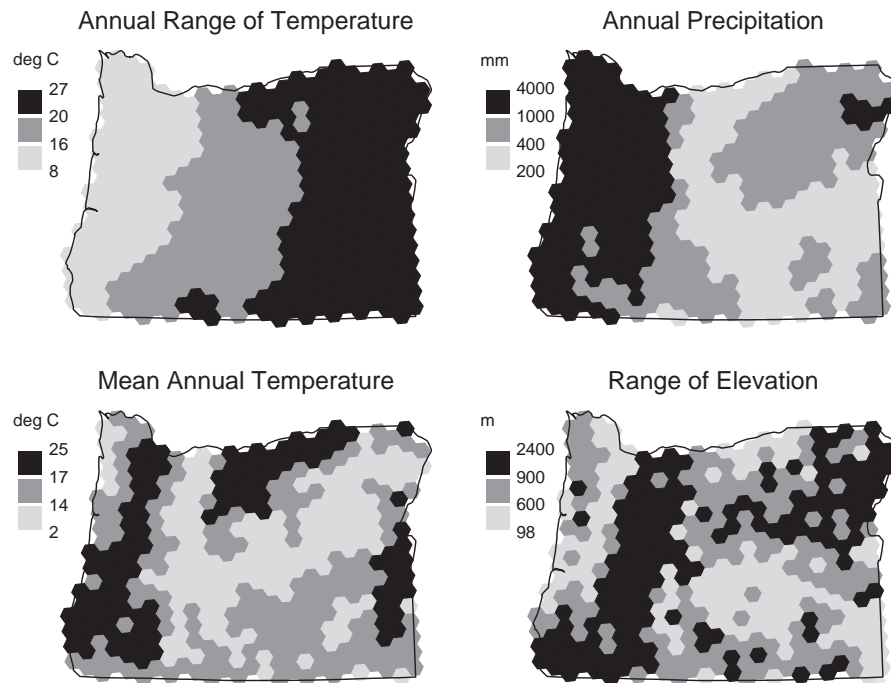


Figure 7. Choropleth maps for conditioning variables.

found; range of elevation is lowest in the valleys and basins, as would be expected.

Despite the evident spatial pattern in mean annual temperature and range of elevation, the conditioning maps, figure 5, show almost no pattern in their relationship with bird-species richness. This might have been expected from the scatter-plots in figure 6 where the splits in the conditioning variables are also displayed. From this evidence we can doubt that the global pattern of increasing richness with increasing temperature and increasing variability in elevation holds in Oregon with these data.

But, with annual range of temperature and annual precipitation, the conditioning maps do reveal a relationship in figure 4. Low range of temperature and high precipitation in the upper left panel correspond with low and medium values of richness. However, as figure 4 shows, the opposite is not the case because in Oregon it is intermediate values of precipitation that are associated with the highest richness. The intermediate effect of precipitation is revealed also in the CCmap where most of the high richness values are in the center row, center, and right columns. Annual range of temperature is a measure of seasonality and many higher values of this conditioning variable are associated with higher, not lower, bird species richness.

These investigations suggest that, in Oregon, the regional patterns are different from the global patterns. For two of the conditioning variables there appears to be little relationship, and for the other two the global pat-

terns are, in part, contradicted. This leads us to speculate that more regional mechanisms are controlling the richness pattern in Oregon. Possible influences could be differences in vegetation structure between eastern and western Oregon forests. The eastern forests have the highest values of species richness and are generally more varied in vertical structure than the western forests. Another possible influence is the intrusion into Oregon in the east and south of ranges of species that are primarily found in the Rocky Mountains and in California, respectively. These range intrusions may have caused increased richness values in some areas. Such speculations from our use of CCmaps again help to foster more detailed investigations.

## CCmaps Design and Data Considerations

It is a challenge to develop graphics that both appeal to many people and draw them toward sophisticated analysis. A simple appearance is often appealing. However, providing too few options and graphics to address some of the inherent scientific complexity in a data set fosters naivety and lack of understanding about the underlying data structure. Trivialized summaries can lead to misinterpretation. On the other hand, providing too many options or too much complexity causes many people to shut the door. Obtaining a balance is a non-trivial challenge.

Our inclusion of QQ plots is perhaps too big a step for some users. Here, we discuss some of our considerations

toward obtaining a balance in the CCmaps design and in data processing for CCmaps.

## Design

The CCmaps design requires compromises. Four of these, considered below, are: the number of panels into which a CCmap should be divided, color schemes for the dependent variable, design of partitioning sliders, and the issue of rounding.

How many panels should be used to partition the regions? Kosslyn (1994) cites research indicating the most people handle comparisons among four items with relative ease. The nine panels in the  $3 \times 3$  layout are substantially more than four. However, ~~most people can handle readily~~ tasks like sorting cards with combinations of three symbol sizes and three shapes into a  $3 \times 3$  layout. The presence of two ordered categories and the crossed layout makes such sorting less complex than dealing with nine unordered categories. The task of partitioning continuous data into three ordered categories using sliders is also easy. The learning curve for CCmaps partitioning is very short. Some information in the continuous variables is lost in the coarse categorization process (into low, medium, and high). Nonetheless the  $3 \times 3$  layout of panels is rich enough to represent interactions and quadratic patterns if present.

From one perspective, the three sliders produce a  $3 \times 3 \times 3$  partitioning of regions. This might be a problem for those favoring two-way tables. However, what appears is a  $3 \times 3$  margin view. All three categories for the dependent variable appear in this margin view as different colors distinguishing the three dependent variable categories.

Some analysts may prefer more than three ordered classes for the dependent variable. Brewer and Pickle (2002) provide research indicating the merits of using quantiles (equal percent categories) and favored the use of five classes. However, CCmaps is a more complex setting, focusing on three variables at once rather than one. Parallel structure with the partitioning sliders suggests that using three categories for the dependent variable is a good starting place.

Alternative panel layouts based on the number of classes of conditioning variables are worth considering. Many people are comfortable working with binary splits and are familiar with  $2 \times 2$  tables, so a  $2 \times 2$  layout is a useful alternative. It has the additional merit of increasing the size of maps. In analogy with familiar chi-square tables, Carr, Wallin, and Carr (2000) illustrate a  $2 \times 2$  layout with margins for rows and columns. This is different use of the  $3 \times 3$  layout where the bottom

right panel is the complete map, and the row and column margins show binary splits for the two conditioning variables. More generally, the use of categorical conditioning variables might motivate the use of other layouts where the constraints of screen size and view complexity are limiting factors.

In terms of color choice, CCmaps starts with three class colors that are readily distinguished hues for most people. The much lighter color for background regions is easy to distinguish. Still, there is a wide range of human preference, so CCmaps provide the option for analysts to select their own colors. Brewer (1997) provides considerable guidance about color choice for maps and other graphics, including color scales for the colorblind. More recently, she developed an instructive web site, [www.colorbrewer.org](http://www.colorbrewer.org), as part of our joint Digital Government research. We recommend this site to those seeking guidance on color choices.

We also made some initial decisions about the scale for the partitioning sliders. As a start, we chose a linear scale for depicting each variable and provided distributional annotation in terms of percent. While the scale is simple to understand, the common occurrence of outliers and skewed distributions makes this a resolution-limiting choice. The number of pixels across a slider determines the number of possible category boundaries. When many pixels are devoted to covering intervals with little or no data, resolution becomes poor in high-density intervals. For some distributions and some audiences, a useful alternative is to provide a log scale option that will pull in large positive values. We wanted to start with large audiences and so avoided the log scale in the initial implementation.

After experimenting with outlier treatment options, we settled on the use of piecewise linear sliders. These simplify to the ordinary linear sliders when there are no outliers. Our approach follows the notion of adjacent values used in Tukey's box plots (see McGill, Tukey, and Larsen 1978). The upper adjacent value is the smallest value less than or equal to the third quartile plus 1.5 times the interquartile range. Any value more extreme is considered an outlier. If there are outliers with high values, we assign 30 pixels of our 440-pixel slider to cover the interval between upper adjacent value and the maximum value. We add the adjacent value to the slider annotation and emphasize the change in scale for this part of the slider by reducing the slider thickness. The thinner bar for large values indicates immediately that there is one or more large-value outliers. The treatment for low-value outliers is analogous. The central part of the slider is linear between adjacent values. Thus a slider may have one, two, or three linear sections. When there

are no outliers, the lower and upper adjacent values are the minimum and maximum values, and the slider becomes the simple linear slider as shown in figures 1 and 8. Occasionally, the most extreme value is not much more extreme than the corresponding adjacent value. In such cases, using thirty pixels to reach the extreme value would reduce the slider resolution in the body of the data, so the slider algorithm reverts to the simple linear form for that end of the distribution. While subject to further fine-tuning, the algorithm, as implemented, successfully protects users from starting with a really bad scale when selecting a new variable with outliers. For those wanting more control, a menu item provides for manual removal or resetting of values used as adjacent values.

In terms of distributions, a more universal approach is to make the scale linear in terms of cumulative percents and to annotate the slider category boundaries with variable values. This is less than ideal for those wanting linearity in the original data units. We like the piecewise linear approach since it often provides visual warning about outliers, but we plan to make an approach that is linear in cumulative percents available as an alternative in a later version of the software.

A fine detail is the treatment of ties at slider boundaries. The slider algorithm assigns regions with values tied with a slider boundary to the interval involving smaller values. To keep regions with tied minimum values from being excluded from the interval with the low values, the slider minimum is made slightly

smaller so there will not be a tie. This treatment allows all of the regions to be assigned to one category and shown in one color. Figure 8 provides an example.

To produce this view, we clicked on the middle panel of figure 1 to enlarge it. We set both the lower and upper inner boundaries in the lung cancer slider at the top to be just below the minimum value. All HSAs are then assigned to a single category, the red category. We then set the precipitation (bottom) slider so that all HSAs are also assigned to one category (the middle category). Next we use the right slider as a filter slider to show the regions with high percent of households below the poverty level. Such simple views can be highly informative and provide a good starting place to introduce CCmaps in presentations.

Adapting Eick's (1993) slider example, Carr, Wallin, and Carr (2000) show a density estimate for the variable in the background of the partitioning slider (interactive sliders were not yet implemented). Such density estimates can help guide partition choices. When a density is bimodal or multimodal, natural choices are to partition between the bumps. We plan to provide access to such density estimates in the future.

A surprisingly difficult choice concerns the display of values as text. Our first prototype produced integers. That version provided for scaling by powers of ten and rounded to integers by default. We chose scale factors by hand so that we would often see two-digit integers and hand set the units in the slider labels accordingly. Work by Ehrenberg (1981) motivated our focus on two-digit

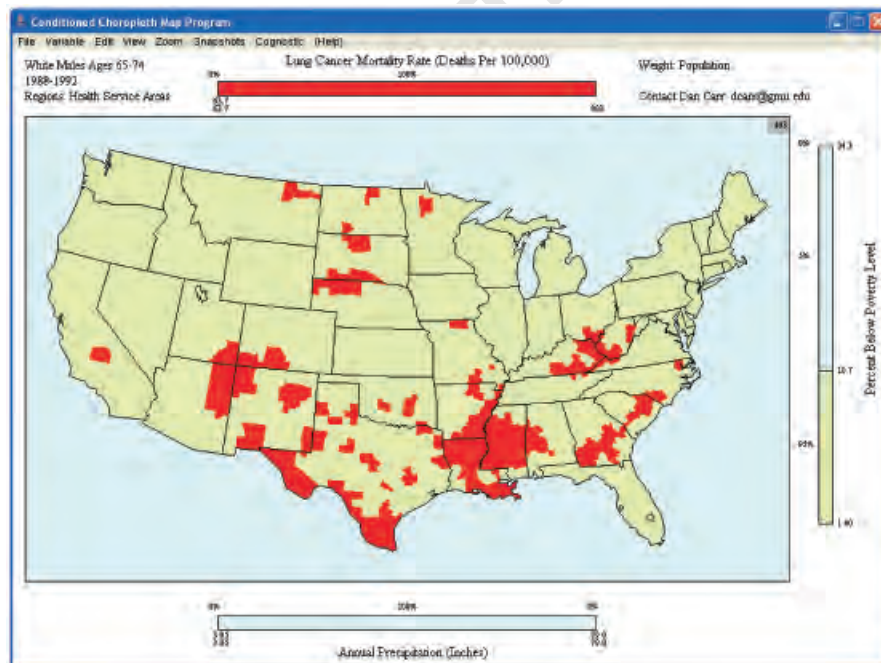


Figure 8. An enlarged CCmap panel with the right slider used as a filter slider.

integers. Conventional wisdom is that people can handle seven plus or minus two digits. Most people working with two two-digit numbers have enough memory left over to store a one- or two-digit answer produced by subtracting or dividing. Showing more digits tends to inhibit thinking quantitatively about relationships. True, people can mentally round before making other computations, but that involves more mental processing, and our working memory is limited.

We changed our approach since many people are uncomfortable with scaling. It is extra work and conflicts with other conventions. The conventional units for mortality rates are deaths per one hundred thousand people. To accommodate such conventions, we began rounding to two significant digits. Showing many digits complicates thinking and often implies an unjustified precision in terms of estimate uncertainty. While the displays looked simple, the lack of feedback when the slider moved was a drawback. Each slider in our  $1024 \times 768$  layout provides 440 options for boundaries, but many boundary labels look the same when rounded to two digits. We now have come to the compromise of rounding the values to three significant digits. Another reasonable approach is to determine the number of digits to show based on the change per pixel. In time, we may add some transformation and display options.

## Patch Maps and Smoothing

John Tukey warns against the use of what he calls “patch maps.” In our interpretation, patch maps include classed choropleth maps showing a single value for each region in a set of political or related enumeration regions (e.g., the HSAs used in our health example). Elections and region taxes are among the few things that respect political boundaries. Mortality rates are not step functions that are constant within political regions. The geometry of the reporting regions constitutes a noise factor that is generally ignored in the attempt to represent the underlying spatial variation of rates. Choropleth maps are thus often ill suited to depicting spatial distributions of continuous mortality rates that vary in places other than region boundaries (MacEachren and DiBiase 1991). We acknowledge this deep weakness of choropleth mapping that underlies CCmaps.

However, maps based on enumeration units are familiar to analysts and provide a common base through which multiple variables can be integrated. The availability of data is often crucial to the identification of multivariate relationships. Due to increasing confidentiality concerns, access to federal statistics for regions smaller than counties is becoming increasingly difficult,

except for some categories of data from the Bureau of the Census. While Tukey recommended prayer to those who use patch maps, he nonetheless was a strong advocate of taking next steps. In our opinion, partitioning regions to help control for the variation due to related variables is a constructive next step, given the constraints on data availability.

Smoothing methods take many forms (see Cressie 1993; Hastie, Tibshirani, and Friedman 2001) and provide an approach to converting region estimates to estimates on a regular grid. Corresponding displays, such as isopleth maps and concomitant 3-D, color-draped surfaces, avoid the visual artifacts of political boundaries. Smoothing borrows information from neighboring regions and so tends to improve estimates. Smoothing, of course, cannot recover spatial information lost through summarization to large regions, when there is substantial spatial variation at a finer scale. Nonetheless smoothing often provides a more realistic view of the underlying phenomena (MacEachren 1994). Plans for future extension to CCmaps include the addition of smoothed map views.

Smoothed views of the relationship between two variables, like those in Figure 2a and 2b, are informative even though spatial context is lost. As indicated by Cleveland, Grosse, and Shu (1992), smoothed scatterplots conditioned by one or two additional variables enable the study of higher dimensional relationships. Since submitting this article for review, we implemented the proposed view of loess-smoothed scatterplots stacked in three vertical panels. One of the conditioning variables used in the map view is shown as a continuous variable on the x-axis. The other conditioning variable remains attached to the right conditioning slider. The right conditioning slider then controls the assignment of regions to scatterplot panels. A selection box controls whether the local smooth is constant, linear, or quadratic, and an additional single parameter slider controls the span of the local smoothing window. This approach brings dynamic control to conditioned, smoothed scatterplots.

An approach of Scott and Whittaker (1996), Whittaker and Scott (1999), and Scott (2000) can address the next step of smoothing that involves both spatial coordinates and attributes. They developed fast smoothing methods for treating cases as neighbors based on both their spatial coordinates and attributes. Their subsequent smoothed and conditioned maps provide a sophisticated and attractive way of looking at multivariate, spatially indexed relations. The smoothing requires large datasets and is still too slow to be directly a part of interactive software. We plan to import the smoothed results.

## Handling More Regions and Slider Variations

As CCmaps addresses examples with more and more regions, issues arise concerning the ability to see small regions. We have added a zoom capability to show the same part of the map in all the panels as illustrated in Carr, Wallin, and Carr (2000). However, we have yet to fully address the option for narrowed focus, that includes computing statistics for only the regions shown and the option for resetting slider scales. An immediate issue is whether or not to include partially shown regions in the highlighting and the statistics. Since regions that have few pixels in the zoomed view may go unnoticed, we are inclined to exclude all regions that cross the zoom boundary and color excluded regions in the background color.

One test version of CCmaps set more extreme scale limits and provided two additional mouse-adjustable boundaries. The adjustable, outer-partition boundaries operated in a manner similar to filter sliders (also called “focusing”; see Haug et al. 1997). We observed little use of this filtering generalization, noted labeling complications, and returned to the simpler approach described above.

We also tested several updating algorithms so the display of regions would keep up with the partitioning sliders. Since the regions always appear, updating the display only necessitates switching the color of the regions. This makes it possible to use real-time, color-look-up-table (CLUT) methods as a fast way to update all the maps when the analyst moves the sliders. However, for our JAVA implementation, we did not want to deal with machine dependencies associated with CLUT methods. With increasingly fast computers, human mouse-based adjustment of a slider boundary tends to be slow compared to computer cycles checking the mouse position. Typically, a modest number of regions change categories in a mouse checking cycle. The simple painting algorithm we chose restricts repainting to the regions that change. We handle the updating of panel statistics similarly. The slider annotation itself is extremely fast since we precompute the slider quantiles and percents, and only pixel-indexed lookup is involved.

In working with roughly eight hundred health service areas, our approach was fast enough for immediate smooth response on the computers we have tried so far. The response time depends upon computer speed, the number of polygons drawn, and the complexity of the polygons. If the polygon boundaries are too detailed, there is repeated drawing on the same pixel, and it may be advantageous to use thinned boundaries. With

roughly three thousand modestly thinned counties boundaries, 2 GHz computers or faster support immediate response except for very fast mouse movements when an intermediate update may be noticed. CCmaps has a zoom option, so there are some applications in which keeping more boundary details is a consideration. While we are still making revisions to improve performance, our motivation is increasingly to help those with slower computers or applications involving many thousands of regions. With graphic card speed doubling every six months, and with increasing ability to exploit this speed, updating speed to keep up with the sliders should soon become a nonissue.

## More Planned Extensions and Closing Remarks

The possible extensions of CCmaps are numerous. Two planned extensions are the development of versions for different languages and the addition of point and line map layers with slider control. Two additional planned extensions need a more detailed description: the display of categorically indexed multivariate data and the addition of cognostics (computer-guided diagnostics) for more strategic interaction. The descriptions are followed by some closing remarks about the access to data and the use of CCmaps.

## Multiple Dependent Variables

One of our high-priority tasks is extending CCmaps to handle multiple dependent variables. As indicated earlier, typical dependent variables are indexed by values of a category, such as sex, or by values of an ordered variable, such as time. The situation is different than that in the current CCmaps because each dependent variable provides a complete map (as proposed by MacEachren 1995, but with the addition of conditioning capabilities). With separate data for males and females and for three time periods, there would be a  $2 \times 3$  layout of complete maps. The presence of complete maps (or matched region maps when one conditioning variable partitions regions) opens the door to the direct display of map differences and/or contrasts. Depicting more categories than three opens the door to rapid selection of categories and motivates development of a different selection widget. Depicting more time periods opens a further door to animation over time of spaced categorical selectors (building on the concept of manipulable, focusable animations proposed in MacEachren et al. 1998).

## Cognostics and Strategic Interaction

John Tukey (1982) coined the term *cognostics*, meaning computer-guiding diagnostics, as an approach to selecting views of data more strategically. If we ask for fifty interesting views, the diagnostic guides the computer in how to rank and select views. While still in the context of having too many potential plots to view, some researchers, such as Carr 1990, have used this term more in the sense of the analyst being guided. The analysts select diagnostics, and the computed results provide guidance to the analysts for view selection. A strategic approach to plot examination then calculates a figure of merit for each view (a cognostic) to both prioritize and limit the examination of plots.

CCmaps poses the problem of too many views. Figure 1 has roughly eight hundred HSAs. When eight hundred regions are ordered by one variable with distinct values, there are combinations of 799 gaps between values, taken two at a time, or 318,801 possible ways to produce three classes with at least one region in each class. The classification possibilities for three variables lead to roughly  $3.2 \times 10^{16}$  different views when using a quick calculation that ignores ties, multivariate constraints, and slider resolution. It is of little comfort that there are only  $9.12 \times 10^{14}$  ways to set the six slider boundaries of the three 440 pixel sliders. Some guidance toward suggestive views seems appropriate.

Since the number of possible views is large, a practical approach is to obtain a cognostic for a subset of views corresponding to a manageable subset of slider settings. One recipe for an eight-hundred-region application is as follows: Sort the regions based on the values of one partitioning variable. Determine nine values that will partition the sorted list into ten classes with roughly eighty regions per class. Repeat this sorting and partitioning process for the second partitioning variable. Then, choose two partition values from each set of nine partitioning values. Using just two partition values for a partitioning variable creates three classes, just like setting two values using CCmaps sliders. There are combinations of nine boundaries taken two at a time or thirty-six ways of doing this for each variable. With two partitioning variables, there are  $36 \times 36 = 1296$  different combinations of choices. The cognostic provides a ranked list of these 1296 views. The interface allows the analyst to instantiate slider settings and views by clicking on items at the top of the list. The top (most interesting) slider settings provide a starting place for further manual exploration. Other slider settings can help provide visual calibration by providing more typical views.

Analysts may consider various cognostics to provide summary statistics based on  $3 \times 3$  partitioning of regions. For mortality rates, one possibility is to use the Mantel-Haenszel statistic to obtain a  $p$  value as a basis for ranking views. Small  $p$  values indicate departures from what one expects if the partitioning variables are independent from the dependent variable. The calculations are simple. With the population and rates at hand for the regions in each cell, determining the total number that died and survived per cell is straightforward. Starting with this summary, the Mantel-Haenszel calculations are only a little different than well-known Chi-squared calculations. Another possibility is to use analysis of variance for the rates with adjustments made for the estimated variance of the rates. Related statistics and corresponding  $p$  values can point to partitions that yield statistically unusual one-way effects or two-way interactions.

There are complications with the above suggestions. First, mortality rates are a special case that can be put back in categorical context, and even that is an approximation when working with age-adjusted rates. Many other dependent variables are not amenable to this approach. Second, the sliders can easily be set so some panels are empty. When some panels have few or no regions, the choice of an analysis of variance model can be controversial and  $p$ -value approximations can become questionable. Third,  $p$ -values may be taken too literally and used in the sense of significance testing without adjusting for the search process.

We implemented the first cognostic in CCmaps since this article was submitted for review. It is a population-weighted variant of the familiar  $R^2$  value used to assess the percent of variability modeled relative to fitting just the mean. The model consists of fitting panel-weighted means to region values in each panel. The algorithm exploits the structure of the statistic and the grid search to reuse calculations. It is very fast, so the first implementation uses fourteen candidate bounds per slider rather than the nine suggested in the recipe above. The number of candidate bounds will be made a parameter for those desiring a more or less detailed grid search. The implementation enables stepping forward or backward through the top-ranked views one mouse click at a time.

Choosing additional cognostics that lead to insight is an area for more research. These can be based on sophisticated regression models. The ranked list of partitions (slider settings) can be generated outside of CCmaps and imported into CCmaps. Cognostics can also provide guidance as to variable selection. Currently, we are running the  $R^2$  algorithm outside of CCmaps for all pairs of conditioning variables to obtain empirical guidance on variable selection.

The cognostics being suggested have some conceptual similarities to projection pursuit algorithms (Cook et al. 1995) to regression trees (White and Sifneos 2002). Projection pursuits are different in that they are typically steepest ascent algorithms designed to find a 2-D projections of many variables whose figure of merit is a local optimum. Common regression trees are different in that partitioning is based on one variable at time rather than two and that each partition is typically a binary partition rather than a  $3 \times 3$  partition. Currently the CCmaps partitioning is not iterated to produce a tree as in regression trees, but that is a possibility for further consideration.

There is much work to do in promoting the informed use of cognostics. It remains a challenge to communicate to potential users of our exploratory tools that by searching enough, we can often find extreme patterns even when only random variation is involved, and we can overfit the data when there is underlying structure. Extreme views do not necessarily promote the generation of good hypotheses, and overfit models do not fit so well when additional data is available for evaluation. The need to incorporate scientific background into the processes of hypothesis generation and critique remains.

### Making Tools Available

Our goal is to put new tools in the hands of a wide range of analysts. We have, as yet, not subjected CCmaps to formal usability assessment since we have been more focused on providing additional capabilities and addressing issues based on our own use of CCmaps. We see opportunities for improvement in CCmaps from boundary file readers to menus and widgets. Early feedback from students and others concerning existing applications is often enthusiastic. As might be expected, the maps are popular, and the QQ plots are not. CCmaps is useful for classroom instruction. It is too soon to comment much about research use of CCmaps. Edward Wegman, in personal communication, reports that his CCmaps presentation on the ecology of alcoholism for zip code regions in Erie County, New York, was met with great enthusiasm. There are many communities that have not seen the dynamic capabilities provided by CCmaps, so early response is likely to be good. More time will be needed to assess the impact of CCmaps on research.

The major thrust of our future CCmaps research is to address increasingly complex scenarios while striving to keep things simple. Some people will use CCmaps to generate misguided hypotheses. Some will fail to distinguish between hypotheses and conclusions. Nonetheless,

we believe that the net effect of providing CCmaps for educational and research purposes will be beneficial. It is not only correct to look at data prior to developing hypotheses; such looking can be instrumental in producing insight into multivariate patterns of geospatially indexed data and related models.

### Acknowledgements

Duncan MacPherson developed the interactive proof of the concept version of CCmaps. The current software is largely the work of Yuguang Zhang and Yaru Li.

The majority of research for this article was supported by National Science Foundation (NSF) collaborative grants #EIA9983461 and #EIA9983451. EPA cooperative grant No. CR 825564-01-0 funded the research producing the initial static designs. Early thoughts about the interactive implementation were promoted by a small contract with NCHS. The article has not been subject to review by NSF or National Center for Health Statistics (NCHS). This document has been subjected to the U.S. Environmental Protection Agency's peer and administrative review and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. The conclusions and opinions are solely those of the authors and are not necessarily those of the agencies.

### References

- Andrienko, G. L., and N. V. Andrienko. 1999. Interactive maps for visual data exploration. *International Journal of Geographic Information Science* 13 (4): 355–74.
- . 2001a. Constructing parallel coordinates plot for problem solving. In *First international symposium on smart graphics*, ed. A. Butz, A. Krueger, P. Oliver, and M. Zhou, 9–14. Hawthorne, NY: ACM Press.
- . 2001b. Interactive cumulative curves as a tool for exploratory classification. In *9th Annual Conference GIS Research in the UK*, ed. D. B. Kidner and G. Higgs, 439–42. Glamorgan, Wales: University of Glamorgan.
- Anselin, L. 1994. Exploratory spatial data analysis and geographic information systems. In *New tools for spatial analysis*, ed. M. Painho, 45–54. Luxembourg: Eurostat.
- Bertin, J. 1983. *Semiology of graphics: Diagrams, networks, maps*. Madison: University of Wisconsin.
- Brewer, C. A. 1997. Spectral schemes: controversial color use on maps. *Cartography and Geographic Information Systems* 24 (4): 203–20.
- Brewer, C. A., and L. Pickle. 2002. Evaluation of methods of classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers* 92:662–81.
- Brown, J. H., and M. V. Lomolino. 1998. *Biogeography*. 2nd ed. Sunderland, MA: Sinauer Associates.

- Brunsdon, C. 2001. The comap: Exploring spatial pattern via conditional distributions. *Computers, Environment and Urban Systems* 25 (1): 53–68.
- Caldecott, J. O., M. D. Jenkins, T. H. Johnson, and B. Groombridge. 1996. Priorities for conserving global species richness and endemism. *Biodiversity and Conservation* 5:699–727.
- Carr, D. B., R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. 1987. Scatterplot matrix techniques for large N. *Journal of the American Statistical Association*, 82:424–36.
- Carr, D. B., A. R. Olsen, S. M. Pierson, and J.-Y. P. Courbois. 2000. Using linked micromap plots to characterize Omernik ecoregions. *Data Mining and Knowledge Discovery* 4:43–67.
- 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* 19 (4): 228–36. 271.
- Carr, D., J. F. Wallin, and A. Carr. 2000. Two new templates for epidemiology applications: linked micromap plots and conditioned choropleth maps. *Statistics in Medicine* 19: 2521–38.
- Cleveland, W. S. 1985. *The elements of graphing data*. Summit, NJ: Hobart Press.
- . 1993. *Visualizing data*. Summit NJ: Hobart Press.
- Cleveland, W. S., E. Grosse, and W. M. Shu. 1992. Local regression models. In *Statistical models in S*, ed. J. M. Chambers and T. J. Hastie, 309–72. Pacific Grove, CA: Wadsworth and Brooks/Cole.
- Cook, D., A. Buja, J. Cabrera, and C. Hurley. 1995. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics* 4:155–72.
- Cook, D., J. Symanzik, J. J. Majure, and N. Cressier. 1997. Dynamic graphics in a GIS: More examples using linked software. *Computers & Geosciences, Special issue on Exploratory Cartographic Visualization* 23 (4): 371–86.
- Cressie, N. A. C. 1993. *Statistics for spatial data*. New York: Wiley.
- Daly, C., R. P. Neilson, and D. L. Phillips. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology* 33:140–58.
- Dykes, J. A. 1997. Exploring spatial data representation with dynamic graphics. *Computers & Geosciences* 23:345–70.
- . 1998. Cartographic visualization: exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv'. *The Statistician* 47 (3): 485–97.
- Eick, S. 1993. *Data visualization sliders*. Naperville, IL: AT&T Bell Laboratories.
- Ehrenberg, A. S. C. 1981. The problem of numeracy. *The American Statistician* 35 (2): 67–71.
- Emerson, J. D., and D. C. Hoaglin. 1983. Analysis of two-way tables by medians. In *Understanding robust and exploratory data analysis*, ed. D. C. Hoaglin, F. Mosteller, and J. Tukey, 16–207. New York: Wiley.
- Harrower, M., A. M. MacEachren, and A. Griffin. 2000. Design, implementation, and assessment of geographic visualization tools to support earth science education. *Cartography & Geographic Information Systems* 27 (4): 279–93.
- Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The elements of statistical learning, data mining, inference, and prediction*. New York: Springer-Verlag.
- Haug, D. et al. 1997. Implementing exploratory spatial data analysis methods for multivariate health statistics, *Proceedings of GIS/LIS '97*. AAG, Cincinnati, OH, 205–12.
- Kosslyn, S. M. 1994. *Elements of graph design*. New York: W. H. Freeman and Company.
- Lawson, A. B. 2001. *Statistical methods in spatial epidemiology*. New York: John Wiley and Sons.
- MacDougall, E. B. 1992. Exploratory analysis, dynamic statistical visualization, and geographic information systems. *Cartography and Geographic Information Systems* 19 (4): 237–46.
- MacEachren, A. M. 1994. SOME truth with maps: A primer on design and symbolization. *AAG Resource Publications in Geography*. Washington, DC: Association of American Geographers.
- . 1995. *How maps work: Representation, visualization and design*. New York: Guilford Press.
- MacEachren, A. M., F. P. Boscoe, D. Haug, and L. W. Pickle. 1998. Geographic visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics, 87–94. *Proceedings, Information Visualization '98*. IEEE Computer Society, Raleigh-Durham, NC, 19–20 October.
- MacEachren, A., X. Dai, F. Hardisty, D. Guo, and G. Lengerich. 2003. Exploring High-D spaces with multiform matrices and small multiples, 31–38. *Proceedings of the international symposium on information visualization*, Seattle, WA, 19–21 October.
- MacEachren, A. M., and D. W. DiBiase. 1991. Animated maps of aggregate data: conceptual and practical problems. *Cartography and Geographic Information Systems* 18 (4): 221–29.
- MacEachren, A. M., M. Wachowicz, D. Haug, R. Edsall, and R. Masters. 1999. Constructing knowledge from multivariate spatiotemporal data: Integrating geographic visualization with knowledge discovery in database methods. *International Journal of Geographic Information Science* 13 (4): 311–34.
- Marks, D. 1990. The sensitivity of potential evapotranspiration to climate change over the continental United States. In *Biospheric feedbacks to climate change: The sensitivity of regional trace gas emissions, evapotranspiration, and energy balance to vegetation redistribution*, ed. H. Gucinski, D. Marks, and D. Turner P IV-1–IV-31. (Report EPA/600/3-90/078). Washington, DC: U.S. Environmental Protection Agency.
- Master, L. 1996. Predicting distributions for vertebrate species: some observations. In *Gap analysis: A landscape approach to biodiversity planning*, ed. J. M. Scott, T. H. Tear, and F. W. Davis, 171–76. Bethesda, MD: American Society for Photogrammetry and Remote Sensing.
- McGill, R., J. W. Tukey, and W. A. Larsen. 1978. Variation of boxplots. *The American Statistician* 32:12–16.
- Monmonier, M. 1989. Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis* 21 (1): 81–84.
- . 1992. Authoring graphics scripts: Experiences and principles. *Cartography and Geographic Information Systems* 19 (4): 247–60.
- National Climatic Data Center. 2004. [www.ncdc.noaa.gov/ol/climate/online/coop-precip.html#FILES](http://www.ncdc.noaa.gov/ol/climate/online/coop-precip.html#FILES) (last accessed, 12 September 2004).
- O'Connor R. J., M. T. Jones, D. White, C. Hunsaker, T. Loveland, B. Jones, and E. Preston. 1996. Spatial partitioning of environmental correlates of avian biodiversity in the conterminous United States. *Biodiversity Letters* 3 (4): 97–110.

- Pickle, L. W., M. Mingle, G. K. Jones, and A. A. White. 1996. *Atlas of United States mortality*. Hyattsville, MD: National Center for Health Statistics.
- Pickle, L. W., and S. Yuchen. 2002. Within-state geographic patterns of health insurance coverage and health risk factors in the United States. *American Journal of Preventive Medicine* 22 (2): 75–83.
- Richardson, S., and C. Monfort. 2000. Ecological correlation studies. In *Spatial epidemiology, methods and applications*, ed. P. Elliot, J. C. Wakefield, N. G. Best, and D. J. Briggs, 205–20. New York: Oxford University Press.
- Scott, D. W. 2000. Multivariate smoothing and visualization. In *Smoothing and regression: Approaches computation and application*, ed. M. G. Schimek, 451–70. New York: Wiley.
- Scott, D. W., and G. Whittaker. 1996. Multivariate applications of the ASH in regression. *Communication in Statistics* 25: 2521–30.
- Tufte, E. R. 1983. *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- . 1990. *Envisioning information*. Cheshire, CT: Graphics Press.
- . 1997. *Visual explanations*. Cheshire, CT: Graphics Press.
- Tukey, J. W. 1976. Statistical mapping: What should not be plotted. In *Proceedings of the 1976 workshop on automated cartography and epidemiology*, 18–26. Arlington, VA: U.S. Department of Health, Education, and Welfare.
- . 1982. Another look at the future: Computing science and statistics. In *Proceedings of the 14th symposium on the Interface*, ed. K. W. Heiner, R. S. Sacher, and J. W. Wilkinson, 2–8. New York: Springer-Verlag.
- Whitaker, G., and D. W. Scott. 1999. Nonparametric regression of analysis of complex surveys and geographic visualization. *Sankhya*, (Series B. Special issue on small-area sampling) 61:202–27.
- White, D., A. J. Kimerling, and W. S. Overton. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. *Cartography and Geographic Information Systems* 19 (1): 5–22.
- White, D., E. M. Preston, K. E. Freemark, and A. R. Kiester. 1999. A hierarchical framework for conserving biodiversity. In *Landscape ecological analysis: Issues and applications*, ed. J. M. Klopatek and R. H. Gardner, 127–53. New York: Springer-Verlag.
- White, D., and J. C. Sifneos. 2002. Regression tree cartography. *Journal of Computational & Graphical Statistics* 11 (3): 600–14.
- Wilk, M. B., and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55:1–17.

Correspondence: ■ ■ ■

# Author Query Form

---

---

Journal      **ANNA**

Article      **09501002**

---

**1. Disk Usage :-**

Disk not provided

Disk used

Additional Work done:

Disk not used

Disk is corrupt

Unknown file format

Virus found

**2. Queries :-**

While preparing this paper/manuscript for typesetting, the following queries have arisen.

Query No.	Description	Author Response
Q1	PLEASE SUPPLY CORRESPONDENCE DETAILS	