A Theory and Implementation of Smooth Conditional Maps *

J. Blair Christian
Dept of Statistics, MS-138
Rice University
Houston, TX 77005-1892
blairc@stat.rice.edu
713-348-2695

David W. Scott
Dept of Statistics, MS-138
Rice University
Houston, TX 77005-1892
scottdw@stat.rice.edu
713-348-6037

Abstract

This article demonstrates the visual power of continuous smoothing and slicing of multivariate maps, through an example relating the rates of screening tests to colon cancer rates. A new theoretical result provides legitimacy and understanding by demonstrating how the conditional maps are related to the usual single smoothed choropleth map of colon cancer rates.

1 Introduction

The study of spatial data has an important role in the mission of governmental entities. The topic has drawn the interest of statisticians (Cressie, 1991) and geographers (MacEachren, 1995), among others. The main presentation tool is the choropleth map, which displays small geographical units by means of color shading (Tobler, 1978; Tufte, 1990). Two outstanding collections of such maps may be found in Pickle et al. (1996) and Brewer and Suchan (2001). Color choice is critical to successful presentation, and Brewer (1999) has described a theory of successful color mapping. Choropleth maps are inherently discontinuous. Smoothing methodologies such as kriging and nonparametric regression can aid visual understanding, but at the cost of some statistical bias. Whittaker and Scott (1994, 1999) have argued that the needs of many decision makers are aided by such a big picture, without insisting on complete fidelity to the raw data.

The challenge in federal mapping is a better understanding of how several variables are related spatially. While we are very successful at mining (1) the spatial distribution of a variable $z(x, y)$ and (2) the changes in the spatial distribution of a variable over time $z(x, y|t)$, we are less successful understanding how two or more spatial variables, $z_1(x, y)$ and $z_2(x, y)$, are related. Commonly, each of the two variables is grouped into two or three bins (hi/lo or hi/med/lo), resulting in a 2×2 or 3×3 table. Choosing a color for each cell in the table, a choropleth map may be constructed which displays the spatial correlation between the variables.

This approach has limitations. The visual impression is not independent of the number of bins. In a corresponding study of histograms, Scott (1992) illustrates how completely different visual

*Research was supported in part by the National Science Foundation grants NSF EIA-9983459 (digital government) and DMS 99-71797 (non-parametric methodology). The authors would like to thank Linda Pickle and Sue Bell for their collaboration and for providing the data, as well as DG collaborators Drs. Carr, MacEachren, and Brewer.

impressions result from different choices of bin width and bin origin. For mapping data, bin width corresponds to aggregation levels (combining small census tracts, for example), while bin origin corresponds to the choice of cut points for each variable. Scott demonstrated that simply averaging several histograms constructed from shifted bins eliminates the visual ambiguity. Thus, smoothing is an important part of the solution in the present context.

Furthermore, there is no unique bivariate color ordering, so that decoding the $k \times k$ choropleth maps requires careful study. Increasing the number of bins exacerbates the problem of choosing a matrix of colors that can be visually decoded. Even collecting a larger amount of data does not solve this fundamental problem.

Finally, the ability to decode color-matrix choropleth maps depends heavily on the actual pattern of spatial correlation. If the variables are highly correlated, then the spatial correlation will clear if spatial distribution of one of the variables is smoothly varying. (Of course, in such a case, little additional information is learned.)

In the case of density data, the problem of understanding the joint behavior of variables and displaying the relationship in a smooth fashion can be solved through the use of the conditional density plot estimated by an advanced nonparametric estimator (Scott, 1992). This approach works with as many as 3–5 variables. The theoretical basis of this success is reviewed in the next section. In the following section, we introduce a new extension to that theoretical basis for regression data. Then we modify the regression approach for mapping data. Finally, an application is presented and future directions are discussed.

2 Smoothing Conditioning of Density Plots

Suppose the variable, z_1 is of primary interest, but that we have a covariate, z_2, whose relationship to z_1 we wish to better understand. The marginal density, $f(z_1)$, reveals information about clusters and other features of possible interest concerning the primary variable. The joint probability density, $f(z_1,z_2)$, reveals how the variables co-vary. This function can be estimated and displayed either in a contour or a perspective plot. However, since z_1 is the variable of primary interest, a more informative function to plot is the conditional density function, $f(z_1|z_2)$, for several values of z_2. This function can easily be computed knowing $f(z_1,z_2)$ by the formula

$$f(z_1|z_2) = f(z_1, z_2) / \int f(z_1, z_2) dz_1 = f(z_1, z_2)/f(z_2)\,.$$

Note that the denominator provides a normalization constant so that $f(z_1|z_2)$ is a proper density, i.e. integrates to 1. Examining $f(z_1|z_2)$ graphically properly focuses attention on the primary variable of interest, while suppressing too much simultaneous information about the secondary variable z_2. Scott (1992) argued that a smooth animation of the conditional density as z_2 varies over an interval is more effective that a frozen plot of $f(z_1, z_2)$.

Is there anything theoretical to justify this view? Recall the fundamental result in probability that $\int f(z_1, z_2) dz_2 = f(z_1)$. In conditional form, this becomes

$$\int f(z_1|z_2) f(z_2) dz_2 = f(z_1)\,.$$

This lends itself to a powerful interpretation of data analysis. We begin our focus on the primary variable, z_1, through its density function, $f(z_1)$. Examining a sequence of slices, $f(z_1|z_2)$, if we

aggregate those slices, *weighted by the marginal density* of z_2, then we recover the marginal density of z_1. Thus we can interpret the conditional density as being the *change* or derivative of the density of z_1 with respect to z_2. The sequence of conditional views reveals how z_1 is distributed as z_2 varies, and when accumulated, we see the complete view of z_1 by itself.

However, Scott noted that when $f(z_1, z_2)$ is a nonparametric estimator, the quality of $f(z_1|z_2)$ deteriorates in regions where z_2 is rare. He argued that instead of viewing the normalized conditional density, $f(z_1|z_2)$, we should view the unnormalized conditional density slice,

$$s(z_1|z_2) = f(z_1, z_2)\,.$$

Thus in regions of low probability, the function $s(z_1|z_2)$ approaches zero. We learn not only how z_1 varies with z_2, but also how frequently the combination occurs in the entire population. Scott proposed a graphical examination of a animation of the plots $s(z_1|z_2)$ over an interval of z_2, with extensions to three, four, and five variables. Observe now that the relationship becomes

$$\int s(z_1|z_2)dz_2 = f(z_1)\,,$$

i.e. the *unweighted* accumulation of the slices gives the marginal density of the variable of interest.

We have described how smooth slices of regression functions can be useful for the same purpose (Whittaker and Scott, 1994; Scott and Whittaker, 1996; Whittaker and Scott, 1999; Scott and Wojciechowski, 2001). Is there a similar theoretical basis for viewing an animation of sliced maps? This question is addressed in the next two sections.

3 Smoothing Conditioning of Regression Plots

Let us consider a single predictor variable, x, and primary response variable, z_1. You might think of a simplified one-dimensional map on a line (rather than a plane). The average of the variable z_1 is to be plotted "spatially," that is, we wish to plot $\bar{z}_1(x)$. One difference between the regression and mapping settings is that multiple z_1 values may be obtained from a single choice of x, whereas in the mapping setting typically only a single value may be measured.

If we know the joint distribution, $f(x, z_1)$, then we can compute the quantity to be mapped by the formula

$$\bar{z}_1(x) = \int z_1 f(z_1|x)dz_1\,,$$

where the conditional density, $f(z_1|x)$, is obtained as in the previous section. As in the density setting, a "map" or plot of $\bar{z}_1(x)$ gives us a picture of the (spatial) regression curve.

Next, let us introduce a second response variable, z_2, of secondary interest. We are interested in the "spatial" distribution of $\bar{z}_1$ at fixed levels of z_2. Assuming we have the joint density, $f(x, z_1, z_2)$, of all three variables, we may compute

$$\bar{z}_1(x, z_2) = \int z_1 f(z_1|x, z_2)dz_1\,,$$

where $f(z_1|x, z_2) = f(x, z_1, z_2)/f(x, z_2)$. We propose to examine a smooth animation of the function $\bar{z}_1(x, z_2)$ over an interval of z_2 values.

Is there a theoretical justification for viewing $\bar{z}_1(x, z_2)$ as the *change* or weighted derivative of the "map" $\bar{z}_1(x)$ with respect to the new variable z_2? It turns out that the unweighted integral of

$\bar{z}_1(x, z_2)$ over z_2 does not return $\bar{z}_1(x)$. Thus we seek to introduce a weight function, $w(x, z_2)$, so that the weighted view of all the maps does return $\bar{z}_1(x)$; that is,

$$\int \bar{z}_1(x, z_2) w(x, z_2) dz_2 = \bar{z}_1(x) \,.$$

Our result is that

$$w(x, z_2) = f(z_2|x) \,,$$

providing nearly as strong a theoretical basis for the slicing of regression surfaces as the slicing of density surfaces. The outline of the proof is short, so we provide it now.

$$\begin{aligned}
\int_{z_2} \bar{z}_1(x, z_2) w(x, z_2) dz_2 &= \int_{z_2} \left[\int_{z_1} z_1 f(z_1|x, z_2) dz_1 \right] w(x, z_2) dz_2 \\
&= \int_{z_2} \int_{z_1} z_1 \frac{f(x, z_1, z_2)}{f(x, z_2)} w(x, z_2) dz_1 dz_2 \\
&= \int_{z_2} \int_{z_1} z_1 \frac{f(x, z_1)}{f(x)} \frac{f(z_2|x, z_1)}{f(z_2|x)} w(x, z_2) dz_1 dz_2 \\
&= \int_{z_1} z_1 \frac{f(x, z_1)}{f(x)} \left[\int_{z_2} \frac{f(z_2|x, z_1)}{f(z_2|x)} w(x, z_2) dz_2 \right] dz_1 \\
&= \int_{z_1} z_1 f(z_1|x) \left[\int_{z_2} f(z_2|x, z_1) dz_2 \right] dz_1 \\
&= \int_{z_1} z_1 f(z_1|x) \, [1] \, dz_1 = \bar{z}_1(x) \,,
\end{aligned}$$

choosing $w(x, z_2) = f(z_2|x)$, and noting that the integral of any (conditional) density equals 1.

4 Extension to Mapping Data

The discussion in the previous section applied to a random design regression problem. The alternative regression model assumes that the x-variable is not random, but selected. For example, in nonparametric regression a common assumption is that the samples are of the form (x_i, y_i), where $x_i = i/n$. Nevertheless, the same algorithms may be applied in either setting. The mapping data are of this second form, where the x_i represent the centroid of a mapping unit.

The entire discussion in the previous section may be extended to the case of two spatial variables. In this case, the estimate of interest is $\bar{z}_1(x, y, z_2)$ and the appropriate weight function turns out to be $w(x, y, z_2) = f(y, z_2|x)$. Thus an animation of the "maps" $\bar{z}_1(x, y, z_2)$ over a range of values of the variable z_2 may be thought of as a weighted derivative of the one-variable map $\bar{z}_1(x, y)$ with respect to the new mapping covariate, z_2.

The particular algorithm we have chosen to use to estimate $f(x, y, z_1, z_2)$ is the averaged shifted histogram (ASH: Scott, 1992), which is a computationally efficient approximation of the well-known but computationally slow kernel estimator. Further discussion of these ideas may be found in Scott and Whittaker (1996) and Scott and Wojciechowski (2001).

5 Applications to SEER Data

We briefly examine the spatial distribution of colon cancer rates as a function of the percentage of adults having a FOBT (fecal blood test) during a two year period; see Pickle and Su (2002)

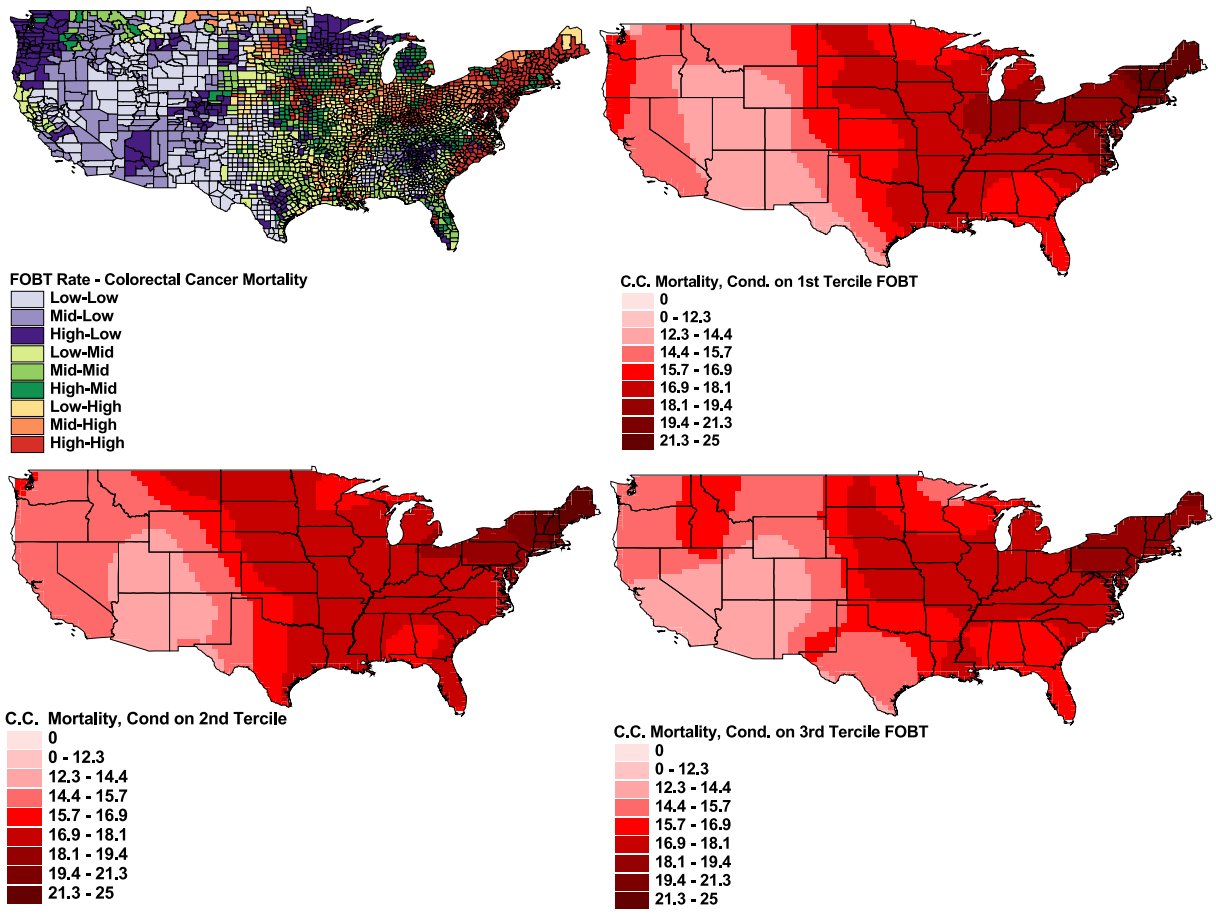

Figure 1: Colon cancer rates versus FOBT rates. One choropleth and 3 ASH estimates.

and SEER (1998). In the Figure above, we display a 3×3 choropleth map, as well as three slices out of an ASH conditional regression map. Clearly, more fine detail is available in the conditional regression maps. Other examples have been truncated due to space limitations here, but will be shown.

6 Discussion and Future Directions

We have provided a new justification for continuous conditional maps as the weighted "derivative" of the unconditional map with respect to a covariate. Weighted integration of the collection of conditional maps results in the original map of the variable of interest. We plan to distribute arcview code of this approach for independent testing and evaluation. We also hope these ideas are implemented by our DG partners. For special situations, an extension to two covariates should be feasible. We hope to find a partner with such data during the coming period.

References

Brewer, Cynthia A. (1999), "Color Use Guidelines for Data Representation," Proceedings of the Section on Statistical Graphics, American Statistical Association, Baltimore, pp. 55-60.

Brewer, Cynthia A., and Trudy A. Suchan (2001), *Mapping Census 2000: The Geography of U.S. Diversity,* CENSR/01-1, U.S. Government Printing Office, Washington DC.

Cressie, Noel A. C. (1991), *Statistics for spatial data,* Wiley-Interscience, New York.

MacEachren, Alan M. (1995), *How maps work,* Guilford Publications.

Pickle, L. W., Mungiole, M., Jones, G. K., and White, A. A. (1996), *Atlas of United States Mortality,* HHS-CDC, Hyattsville, MD.

Pickle, L. W. and Su (2002), "Within-state geographic patterns of health insurance coverage and health risk factors in the United States," Am J Prev Med 22(2):75-83.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization,* John Wiley, New York.

Scott, D.W. (2000), "Multidimensional Smoothing and Visualization," In *Smoothing and Regression. Approaches, Computation and Application,* M. G. Schimek, Ed., John Wiley, New York, pp. 451-470 (with 5 color plates).

Scott, D.W. (2001), "Parametric Statistical Modeling by Minimum Integrated Square Error," Technometrics, 43, pp. 274–285.

Scott, D.W. and Whittaker, G. (1996), "Multivariate Applications of the ASH in Regression," Communications in Statistics, 25, pp. 2521-2530.

Scott, D.W. and Wojciechowski, W.C. (2001), "Conditioning Multiple Maps," *Proceedings of the First National Conference on Digital Government*, E Hovy, Ed., Digital Government Research Center, Los Angeles, pp. 33-39.

SEER (Surveillance, Epidemiology, and End Results) Program Public-Use Data (1973-1998), National Cancer Institute, DCCPS, Surveillance Research Program.

Tobler, W. R. (1978), "Data Structures for Cartographic Analysis and Display", Proceedings of the Computer Science and Statistics 11th Annual Symposium on the Interface, pp. 134-139.

Tufte, Edward R. (1990), *Envisioning information,* Graphics Press, Cheshire, CT.

Whittaker, G. and Scott, D.W. (1994), "Spatial Estimation and Presentation of Regression Surfaces in Several Variables Via the Averaged Shifted Histogram," Computing Science and Statistics, 26, pp. 8-17.

Whittaker, G. and Scott, D.W. (1999), "Nonparametric Regression for Analysis of Complex Surveys and Geographic Visualization," Sankhya, Series B, special issue on small-area sampling, P. Lahiri and M. Ghost, Eds., 61, pp. 202-227.