Conditioning Multiple Maps *

David W. Scott
Department of Statistics
Rice University
6100 Main Street
Houston, TX 77005
scottdw@stat.rice.edu

William C. Wojciechowski
Department of Statistics
Rice University
6100 Main Street
Houston, TX 77005
williamc@stat.rice.edu

Abstract

Maps can be used to display the relationship between location and a response variable. In this basic form, there is one map to be displayed. Often times there are additional factors that could affect the response variable. In this case, it is of interest to simultaneously view the spatial relationships and the dependence of the response variable on the additional factors. Because the additional factors will be multi-valued, the number of maps to display can be very large. This large number of maps presents a challenge for effectively presenting the information on the discrete plane of the computer screen. We propose a method that addresses this problem. The method draws one map whose display depends on the values of one or more factors (conditioning variables). As the values of the conditioning variables change, the information displayed on the map is calibrated in real-time.

1 Introduction

The conditional density function accurately represents the variation of one variable, z_1, given the value of a second variable, z_2. This conditional density,

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

may be explored graphically or algebraically. If a parametric form is not known for the joint density $f(z_1, z_2)$, from which $f(z_2|z_1)$ is derived, then nonparametric multivariate density estimation (Scott, 1992) may be used to accurately estimate the density.

Spatial data present some interesting opportunities and challenges. Here, the variables z_1 and z_2 vary over space (x, y) and may be re-written as $z_1(x, y)$ and $z_2(x, y)$. Often, these data are collected on an irregular mesh (such as county boundaries) and are presented as a choropleth map (Tufte, 1990; Tobler, 1978). Choropleth maps are constant over the counties, and hence discontinuous between counties. The result can be a map that misrepresents the underlying spatial process. A more accurate approach would be to smooth the data.

*Research was supported in part by the National Science Foundation grants NSF EIA-9983459 (digital government) and DMS 99-71797 (non-parametric methodology).

Two issues arise. First, when may one smooth such data? Certainly the choropleth map is accurate, but if in fact the data values are subject to measurement error, or if the underlying process may be usefully viewed as continuous, then smoothing may be appropriate. A nonparametric smooth is provided by the Nadaraya-Watson estimator (Watson, 1964; Altman, 1993). Such an estimator is similar to kriging procedures, but does not invoke a number of spatial process assumptions.

The second and more challenging question is the simultaneous consideration of $z_1(x, y)$ and $z_2(x, y)$. For example, in the Atlas of United States Mortality (Pickle et al 1996), an appendix is provided on pages 208-209 of "correlate variables by county" (poverty level, college education, hispanic origin, and urbanization level). One might imagine that certain cancers are related to these variables, but as separate maps, the "correlation" is left to the sophisticated reader. Numerous attempts to use two aspects of display (such as hue and saturation, or color gridding) to indicate simultaneous levels of z_1 and z_2 are often surprisingly difficult to "read". For example, cancer incidence might be grouped into terciles, and poverty rates grouped into terciles, giving 9 "simultaneous conditions" that can be represented on a choropleth map. A "smooth" map would be interesting, provide more detail, and give a broader view of the "correlation". Here we discuss one such mapping technique based upon the averaged shifted histogram, and we illustrate its use. Our basic idea is to extend and combine the estimates $z_1(x, y)$ and $z_2(x, y)$ into a single multivariate quantity $z_1(x, y, z_2)$ which can be displayed for several values of z_2 over a regular mapping surface. The utility of this idea is explored below.

2 Averaged Shifted Histogram

A basic and familiar graphic for displaying the probability density function of continuous data is the histogram. The appearance of a histogram is determined by three items: the data, bin-width, and origin. Because the ASH is the average of many histograms, each with a different origin, it removes the need to select a starting point. The averaged histograms use the same data and bin-width. The basics of the ASH as a density estimator and regressor are covered below. The theoretical and computational properties of this estimator have also been studied (Scott 1992). We chose to use the ASH because it is both computationally and statistically efficient.

2.1 ASH Density Estimation

Let $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m$ be a set of m histograms, each having bin width h. The origin of $\hat{f}_i$ is equal to $(i-1)h/m$. By convention, the initial bin origin is equal to 0. The most basic definition of the ASH is given by

$$\hat{f}_{ASH}(x) = \frac{1}{m}\sum_{i=1}^{n} \hat{f}_i(x).$$

Because the ASH is piecewise constant over the intervals $[k\delta, (k+1)\delta]$, it is possible to generalize the ASH. In the above notation, $\delta \equiv h/m$. The interval above is defined to be B_k, the kth bin, and has height, or bin count, ν_k. Using these definitions, the height of the ASH in B_0 is an average of the heights from m shifted histograms. Therefore, the height of bin B_0 is

$$\frac{\nu_{1-m} + \ldots + \nu_0}{nh},$$

where n is the number of data points. Following from this, a general form for the ASH is

$$\hat{f}(x;m) = \frac{1}{nh}\sum_{i=1-m}^{m-1}(1-\frac{|i|}{m})\nu_{k+1} \quad \text{for } x \in B_k \ .$$

If $(1-\frac{|i|}{m})$ is regarded as a weight, then the general ASH can use arbitrary weights. In this case, the ASH is calculated using

$$\hat{f}(x;m) = \frac{1}{nh}\sum_{i=1-m}^{m-1} w_m(i)\nu_{k+1} \quad \text{for } x \in B_k \ .$$

In order that the ASH integrates to 1, the $w_m(i)$ must sum to m. A simple way to achieve this is to define the weights as

$$w_m(i) = m\frac{K(i/m)}{\sum_{j=1-m}^{m-1} K(j/m)},$$

where K is a continuous function defined on $(-1,1)$.

2.2 ASH Regression

Non-parametric regression, or smoothing (Scott, 2000; Scott and Whittaker, 1996), has seen a large amount of attention in statistical literature. These smoothing techniques have also been applied to spatial data (Whittaker and Scott, 1994; Whittaker and Scott, 1999). We propose to expand this work by implementing animated maps instead of static maps. The computational and statistical properties of the ASH make it ideal for spatial data.

Because the ASH is usually associated with density estimation, the regression case will be formulated as a density estimation problem. In particular, the conditional density of z given another variable x is of interest. This conditional density can be written as

$$f(z|x) = \frac{f(z,x)}{f(x)}.$$

In a regression setting, the expected value of Z given X ($E[Z|X]$) is to be predicted. Using the conditional density formula, the conditional expected value is

$$r(x) = E[Z|X=x] = \int zf(z|x)dz = \frac{\int zf(z,x)dz}{f(x)}.$$

To formulate an ASH estimator of $r(x)$, additional notation is necessary. Let h_x be the bin width and m_x be the number of shifts. Then, using h_x and m_x, define $\delta_x = h_x/m_x$. Let $x_1,\ldots,x_n$ represent the midpoints on a mesh separated by a distance of δ_x. Let ν_j represent the number of data points falling in bin j. The weights are given by

$$w_a = \frac{K(a/m_x)}{\sum_a K(a/m_x)},$$

where $-m_x < a < m_x$. Let $\overline{z}_j$ be the average of the z values that correspond to all data points x falling in bin j.

Using the above notation, the ASH estimate of $E[Z|X=x]$ is calculated by

$$\hat{r}(x) = \frac{\sum_a w_a\nu_{j+a}\overline{z}_{j+a}}{\sum_a w_a\nu_{j+a}} \quad \text{for } x \in B_j \ .$$

This notation is easily extended to a two dimensional grid representing coordinates on a map. Furthermore, the extension to a grid of any number of dimensions is straight forward. Because the grid can be in any dimension, the ASH is not limited to using only location to predict a response. Additional covariates, such as time or demographic data, can be used in conjunction with location. Theoretical results and algorithms for the ASH regressor have been developed (Scott, 1992).

3 Mapping Challenges

Often times theoretical results consider cases that are much less complex than those encountered in applications. The challenges presented by map data are not an exception. In the proceeding sections, some of these challenges are described. At this time, the optimal solutions to these challenges are not known. These open problems are areas for interesting research.

3.1 Conditioning

From the previous sections, it was seen that regression can be formulated as a conditional density estimation problem. The x and y location variables are two conditioning variables. These conditioning variables are different in the sense that their values will always be visualized on the map display. However, when other conditioning variables (covariates) are considered, how should one display the map to accurately represent the value of the conditioning variables and the response?

There are several known methods for implementing conditioning with statistical graphics. The most basic is slicing. Slicing considers a single value of the conditioning variable. The graphic is then drawn keeping the conditioning variable fixed at that value. Although this technique can be useful, there are drawbacks to its use. Because conditioning variables are often continuous, there are an infinite number of graphics that can be drawn. Which of these infinite possibilities should be displayed? By using animation, a finite number of graphics can be displayed. Animation is achieved in the following manner. As the value of the conditioning variable changes, the graphic is updated to reflect the current value of the conditioning variable. If the changes are updated in real-time, an animation effect will be produced.

Another conditioning technique is coplots (Cleveland, 1993). Instead of a single value of the conditioning variable, Cleveland considers a subset. The observations that fall within this subset are used to create the graphic. Several subsets that span the range of the conditioning variable are considered. A graphic is drawn for each of these subsets. How are the subsets selected? The equal-count algorithm (Cleveland, 1993) is one method for creating the subsets. For one conditioning variable, the equal-count algorithm creates subsets that are overlapping line segments. The union of the line segments spans the range of the conditioning variable. The subsets are overlapping hyper-boxes for many conditioning variables. Because coplots creates a finite number of graphics, it possible to create static graphics that span the range of a conditioning variable. The static nature of coplots has two attractive properties. First, a static graphic can be printed. Second, because the entire range of the conditioning variable is visualized in a single view, one does not have to rely on memory to make comparisons.

A technique that combines animation and coplots is dynamic coplots (Wojciechowski and Harner, 1995). This method combines coplots and animation, receiving the benefits of both methods. A finite number of subsets are calculated by the equal-count algorithm. As the subsets are cycled through, the graphic is updated to reflect the current subset value. This cycling creates an animation effect across a finite number of graphics.

Another technique that is a generalization of coplots and dynamic coplots is continous conditioning (Wojciechowski and Scott, 2000). Similar to dynamic coplots, this technique combines coplots with animation. However, continuous conditioning generalizes the conditioning subset. Instead of a line segment or hyper-box, continuous conditioning allows the subset to be of any form. Thus, the conditioning subset could be a point, ellipsoid, or a union of disjoint sets. The distances the observations are from the subset determine the appearance of the graphic.

Although these conditioning methods have been used for statistical graphics, such as the scatterplot. Their use in mapping applications is limited. Therefore, there is an open frontier for applying these conditioning methods to displaying spatial relationships on maps. This is an opportunity for inventive and exciting research.

3.2 Border Constraints

Often times in statistics, observations are assumed to take on values within a well-defined range. Although, borders on maps can be accurately determined, they are not the typical types of boundaries found in statistics. Not accounting for boundaries can result in a less accurate representation of the data. Although, irregular boundaries has been studied for kernel density estimation (Staniswallis, Messer, and Finston, 1990), it is not clear how to implement a solution that is both statistically and computationally efficient for an animated mapping application.

3.3 Aggregated Data

Related to border constraints is aggregated data. Often times map data is available only at certain levels of resolution. The resolution could be at the country, state, county, or census tract level. In this case, there would be one measurement for each unit. This poses an interesting question. Where is the data located? An obvious answer is to place the data points at the centers of the polygons representing the units. However, this can clearly misrepresent the data. As an example, consider population on a state level. It would not be accurate to place New York's population at the center. This creates two scenarios that are open for investigation. First, if no extra information is available, is it optimal to place the data point at the center? Second, if extra related information is available, how can this information be used to more accurately place the data point?

3.4 Performance

Because the goal is to produce animated maps, performance is a critical issue. To begin with, the ASH produces a grid across the range of the location space. It is possible that a large portion of the location space contains bins without data. To save on computational burden, it is desirable to eliminate these bins from being drawn or stored in memory. Because the ASH algorithm allows an empty bin to be easily detected, there is no need to add an additional procedure that searches and eliminates empty bins.

With using an off-the-shelf geographic information system, peformance of the software is an issue. Currently, the ASH has been implemented for predicting a response only using location. The off-the-shelf software is relatively slow at producing a static graphic. The speed is not adequate for animation and either the off-the-shelf software will have to be extended, or a customized solution will have to be produced. These options are currently being pursued.

4 Summary

Creating a map that accurately represents the spatial relationships and dependence on other covariates is a challenging task. These challenges invoke several interesting topics to be studied. One of these is the optimal method for accounting for borders. Although, some work has been done in this area, it is not known how to implement this in an efficient and accurate manner for animating maps. Another area is aggregated data. Where should the data points be place when the data corresponds to an area and not to a single location? A third issue is computation. Because the ASH relies on binning, it has performance advantages for spatial data. Furthermore, the ASH is also statistically efficient, resulting in a relatively accurate estimate. Even with these advantages, the computational burden of producing an animation, will require some in-house software to be developed. Creating animations based on other covariates also presents several interesting areas of study. Exploring different methods of conditioning is anticipated to provide interesting results for mapping applications.

5 Future Directions

The initial stage of this project explored different options for smoothing maps and implemented a basic demo using off-the-shelf software. This demo created a static map that has been smoothed using the ASH. The next obvious step is to implement an animated map that allows additional covariates to be used for prediction. Because the off-the-shelf software had relatively slow performance for the static maps, it is necessary to explore other options for creating the animations. Developing conditioning techniques will be a part of implementing the animation.

To produce "good" graphics, one inherently assumes that the data is "good". If the quality of the data is in question, so is the message delivered by the graphic. Therefore, statistical methods that perform well even when the data contains anomalies are necessary. Robust methods have this type of stability and are a solution to this problem. This problem is recognized and work has been done on two general robust methods. One of these is the L_2E estimator (Scott, 2001; Wojciechowski and Scott, 1999). Another method is a simulation technique (Wojciechowski, 2001). These methods are useful for two separate, but related, purposes. First, it is desirable to reduce the effect of anomalous observations. Second, it is of interest to identify observations that are outlying. For instance, identify which regions of the country have relatively large cancer rates. It is anticipated that these methods will pre-process the data in the background.

References

Altman, N. S. (1993), "Estimating Error Correlation in Nonparameteric Regression," Statistics and Probability Letters, 18, pp. 213-218.

Cleveland, W. S. (1993), *Visulizing Data,* Hobart Press.

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

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

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

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, in press, August.

Staniswalis, J. G., Messer, K., and Finston D. R. (1990), "Kernel Estimators for Multivariate Smoothing," Technical Report 90-01, Biostatistics, Virginia Commonwealth University.

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.

Watson, G.S. (1964), "Smooth Regression Analysis," Sankhya, Series A, 26, pp. 359-372.

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.

Wojciechowski, W.C. and Harner E. J. (1995), "Dynamic Coplots," Computer Science and Statistics, 27, pp. 274-278.

Wojciechowski, W.C. and Scott, D.W. (1999), "Robust Location Estimation with L2 Distance," Computing Science and Statistics, 31, pp. 292-295.

Wojciechowski, W.C. and Scott, D.W. (2000), "High-Dimensional Visualization Using Continuous Conditioning," Computing Science and Statistics, to appear.

Wojciechowski, W.C. (2001) "Robust Modeling", Unpublished thesis, D. W. Scott, Advisor, Rice University.