Multivariate Applications of the ASH in Regression

David W. Scott* — Department of Statistics, Rice University, Houston, TX 77251-1892

Gerald Whittaker — ERS/USDA Room 937, 1301 New York Ave., NW, Washington, D.C. 20005-4788

September 21, 1995

Abstract

A simple algorithm for estimating the regression function over the United States is introduced. The approach allows for data obtained from a complicated sampling design, as well as for the inclusion of a few additional covariates. The regression estimates are obtained from an associated probability density estimate, namely the averaged shifted histogram. The algorithm has proven especially successful over a large mesh, say 300 by 200 nodes, in a data rich setting, even on a 486 computer running Splus. We currently run much higher resolution meshes on a Pentium. Commonly available alternative codes including kriging failed to produce useful estimates in this setting.

1 Introduction

The problem of nonparametric regression has attracted a wealth of attention since the pioneering papers of Nadaraya (1964) and Watson (1964); see Eubank (1988) and Härdle (1990). Available algorithms range from the simple running median, to variational formulations giving rise to spline estimates, to kernel estimates, and finally local polynomial fitting. There has been a great deal of recent discussion about the right and wrong way to do nonparametric regression. Some have argued for the elegance of splines, while others find the local polynomial approach compelling, but some argue for one's personal preference.

From our experience in the density estimation setting, we find that direct methods work well in 1 to 5 dimensions, but even in 3–5 dimensions, the size of the meshes is growing exponentially, and sufficient data often aren't available. In the regression setting, we find that the discussion in the literature has focused too heavily on relatively small 1 and 2 dimensional data sets where most methods perform reasonably well. In this manuscript, we consider a more realistic and stressful problem dealing with farm data such as that routinely surveyed by the U.S.D.A. These surveys result

*Supported in part by NSF and NSA grants DMS-9306658 and MOD 9086-93, respectively. This paper was presented in person by the first author at the IMS meeting honoring the 70th birthday of Professor Robert V. Hogg.

in very large databases over nonuniform spatial meshes (see Figure 1), complicated by nonuniform weighting schemes as well as interest in several covariates.

Large data sets and/or large mesh sizes result in practical problems. Too many regression methods have solutions or algorithms whose exact form is determined by the number of data points (splines, kernels, etc.) that make computation infeasible even on 486 level computers. The key to computational efficiency is the same as for density estimation: **binning** the multivariate data (Scott, 1992; Härdle and Scott, 1992; Fan and Marron, 1994).

Beyond 4 or 5 dimensions, direct mesh methods of any kind encounter practical difficulties resulting from the curse of dimensionality. Some form of advanced projection technology or additive modeling has proven useful (Hastie and Tibsharani, 1990).

However, "real data" can throw a curve at the best planned evaluation of even carefully constructed algorithms. We have mentioned the special problem of large samples. Here we would like to focus on problems resulting from a mixture of spatial and continuous variables. They are: (1) irregular boundary definition, (2) data collected by a sampling design, and (3) a very large mesh required to have high spatial resolution. In principle, an exact irregular boundary scheme can be handled (perhaps with great programming effort), and weighting can be introduced into the estimation phase. However, many simple-minded implementations run into numerical instabilities with large meshes.

We wish to show how simple the binned methods (specifically the ASH or WARP algorithms) can be modified to handle such data, even with very fine 300×200 spatial meshes, on a 486 level machine.

We find that the common focus on boundary behavior is only a minor part of our thinking. Firstly, we are dealing with large samples and thus only a relatively small bandwidth is required. (By way of contrast, many simulation examples involve $n = 100$ 1-dimensional data where the bandwidth may span 1/4–1/2 of the data interval, making boundary conditions dominant.) Secondly, for mapping purposes, we find the boundary effects and corrections of little practical importance towards understanding and summarizing our data exploration/presentation efforts.

Ironically, we have found "internal boundary" situations more of a practical nuisance. These occur in areas internal to the USA, say, where there are no data (because there is no agriculture), inducing a boundary effect caused by sparseness rather than a physical external boundary. We identify this situation by observing how low the density falls in each region where we are evaluating the regression function (i.e., how close to 0 is the denominator?). This is a multivariate version of the well-known practical problem of "extrapolation" of regression estimates beyond the support of the data.

In our experience, many off-the-shelf kriging or regression programs cannot handle large rectangular meshes of 300×200 points covering a mercator projection of the lower 48 states. Rewriting such codes is always a possibility, but we have found that the simple ASH ideas provide excellent estimates and dramatic correlation with actual photographic evidence. Carr (1990) has used raw (hexagonal histogram) bivariate binning techniques. We are interested in providing some additional smoothing (that will provide improved estimation quality) as well as handling additional covariates.

2 Algorithm Motivation

We start with a simple description of the ideas and algorithms for handling (x, y, z) data where (x, y) represents the center of one of our bivariate bins (approximately 10 miles by 10 miles) containing one or more U.S.D.A. sampling units. The variable z represents the quantity of interest; for example, total farm income or the fraction of Federal dollars in farm income. We seek to estimate $\mathrm{E}\left[Z(x, y)\right]$

or $\bar{z}(x,y)$ in areas where $f(x,y) > 0$.

2.1 Kernel Regression Estimation

Let K be a symmetric kernel function with support on $(-1,1)$ satisfying $\int_{-1}^{1} K(t)\,dt = 1$. Given a positive smoothing parameter h, define the scaled kernel function by $K_h(t) = h^{-1}\,K\left(h^{-1}t\right)$. We take as a starting point the well-known result that the Nadaraya-Watson bivariate regression estimator

$$\hat{m}(x,y) = \frac{\sum_{i=1}^{n} z_i K_h(x-x_i) K_h(y-y_i)}{\sum_{i=1}^{n} K_h(x-x_i) K_h(y-y_i)}$$

is the *exact* result of the computation

$$\hat{m}(x,y) = \int_z z\hat{f}(z|x,y)\,dz = \frac{\int z\hat{f}(x,y,z)\,dz}{\int \hat{f}(x,y,z)\,dz}$$

where the trivariate product kernel density estimator is given by

$$\hat{f}(x,y,z) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)\,K_h(y-y_i)\,K_h(z-z_i)\,.$$

Clearly

$$\int \hat{f}(x,y,z)\,dz = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)\,K_h(y-y_i)$$

since $\int K_h(z-z_i)\,dz = \int K_h(z)\,dz = 1$. Also, $\int z\hat{f}(x,y,z)\,dz = \sum z_i K_h(x-x_i)\,K_h(y-y_i)$, since

$$\int zK_h(z-z_i)\,dz = \int (z+z_i)K_h(z)\,dz = 0 + z_i\,,$$

recalling that $\int zK_h(z)\,dz = 0$ (by symmetry).

Clearly, different smoothing parameters h_x, h_y, h_z could be chosen for each dimension. Interestingly, the particular choice of h_z has no effect on the regression estimate!

It is well-known (Härdle, 1990) that local polynomial regression (LPR) estimators and spline methods have equivalent kernel forms. LPR does have the advantage that the kernel adjusts properly at the boundary to reduce bias (Fan, 1992).

However, the practical gain of the bias correction is often small, as $f(x) \to 0$ near the boundary and/or $m(x) \to 0$ near the boundary. Many authors consider only cases where $f(x)$ is nearly constant over a finite interval, or even the simplest case of a fixed equally-spaced mesh. These situations tend to accentuate boundary concerns and problems.

2.2 ASH Density Algorithm

We mimic the simple Nadaraya-Watson idea except on a more computationally oriented estimator, the averaged shifted histogram (ASH), introduced by Scott (1983, 1985, 1992). We remotivate the multivariate ASH.

Let us slightly alter our notation so that

$$x_1, x_2, \ldots, x_{n_x} \qquad y_1, y_2, \ldots, y_{n_y} \qquad z_1, z_2, \ldots, z_{n_z}$$

are the midpoints along each axis of a trivariate mesh of size $n_x \times n_y \times n_z$ with spacings $\delta_x, \delta_y, \delta_z$. Thus

$$\Delta x_i = \delta_x = \frac{h_x}{m_x} \quad \Delta y_i = \delta_y = \frac{h_y}{m_y} \quad \Delta z_i = \delta_z = \frac{h_z}{m_z}$$

for some integers m_x, m_y, m_z and smoothing parameters h_x, h_y, h_z.

Let ν_{jkl} denote the number of data points $(x, y, z)_i$ falling in bin B_{jkl}. Note that $\sum \nu_{jkl} = n$, and we expect many of the ν_{jkl} to be 0.

The "naive ASH" is constructed by "computing" $m_x \times m_y \times m_z$ (different) trivariate histograms, each with rectangular bin size $h_x \times h_y \times h_z$, but with origins shifted by multiples of $\delta_x, \delta_y, \delta_z$ along the coordinate axes. To be specific, one bin is anchored at the point $(j\delta_x, k\delta_y, l\delta_z)$, as j, k, l each range from 0 to $n_x - 1, n_y - 1, n_z - 1$.

Scott (1985) showed that this was a special case of a general weighting scheme:

$$\hat{f}_{jkl} = \hat{f}(x_j, y_k, z_l) = \frac{1}{n\delta_x\delta_y\delta_z} \sum_{a,b,c} w_{abc}\nu_{j+a,k+b,l+c}$$

where the sums range over $-m_x < a < m_x$, $-m_y < b < m_y$, and $-m_z < c < m_z$, and

$$w_{abc} = \frac{K\left(\frac{a}{m_x}\right) K\left(\frac{b}{m_y}\right) K\left(\frac{c}{m_z}\right)}{\sum_a \sum_b \sum_c K\left(\frac{a}{m_x}\right) K\left(\frac{b}{m_y}\right) K\left(\frac{c}{m_z}\right)},$$

where K is supported on $(-1, 1)$ as before. Note that in an obvious notation, $w_{abc} = w_a w_b w_c$. This is a classic discretization scheme. The weights $\{w_a, w_b, w_c\}$ need only be computed once.

We first verify that the trivariate ASH is indeed a density function. Clearly it is nonnegative and may be shown to integrate to 1.

In practice, the array $\{\hat{f}_{jkl}\}$ is initialized to all 0's, and then the influence of every bin B_{jkl} for which $\nu_{jkl} > 0$ is added to the appropriate subset of $\hat{f}_{jkl}$.

We could define $\hat{f}(x, y, z)$ to be a spline surface interpolated from the above array, but for simplicity, we take it to be constant over each bin B_{jkl} and assume it vanishes outside the mesh; that is, $\hat{f}(x, y, z) = 0$ there.

2.3 ASH Regression Algorithm

Following the Nadaraya-Watson motivation, the ASH regression estimator is found by computing

$$\begin{aligned}\hat{m}_{jk} &= \hat{m}(x_j, y_k) = \mathrm{E}\left(Z | X = x_j, Y = y_k\right) \\ &= \int z\hat{f}(z|x_j, y_k)\, dz = \frac{\int z\hat{f}(x_j, y_k, z)\, dz}{\hat{f}(x_j, y_k)}.\end{aligned}$$

The numerator can be computed by integrating bin by bin along the z axis:

$$\sum_{l=1}^{n_z} \int_{z_l - \delta_z/2}^{z_l + \delta_z/2} z\hat{f}(x_j, y_k, z = z_l)\, dz = \sum_{l=1}^{n_z} \delta_z z_l \hat{f}(x_j, y_k, z_l),$$

since $\int z\, dz = \delta_z z_l$ for the limits given (recall $\hat{f}$ is constant over each bin). Thus

$$\hat{m}_{jk} = \frac{\sum_l \delta_z z_l \left\{\frac{1}{n\delta_z\delta_y\delta_z} \sum_a \sum_b \sum_c w_{abc}\nu_{j+a,k+b,l+c}\right\}}{\frac{1}{n\delta_x\delta_y} \sum_a \sum_b w_{ab}\nu_{j+a,k+b}}$$

$$= \frac{\sum_a \sum_b w_{ab} \sum_c w_c \sum_{l=1}^{n_z} z_l \nu_{j+a,k+b,l+c}}{\sum_a \sum_b w_{ab} \nu_{j+a,k+b}} .$$

Now the final sum in the numerator can be computed by observing that it is almost a conditional expectation:

$$\sum_{l=1}^{n_z} z_l \frac{\nu_{j+a,k+b,l+c}}{\nu_{j+a,k+b}} \cdot \nu_{j+a,k+b} = \bar{z}_{j+a,k+b}\, \nu_{j+a,k+b}$$

as we let $m_z \to \infty$ (or equivalently let $\delta_z \to 0$ with h_z fixed), where

$$\bar{z}_{ab} = \frac{1}{n_{ab}} \sum_{(x,y,z)_i \in B_{ab}} z_i .$$

Continuing, we note that $\sum w_c = 1$, so that we finally arrive at the final form of the ASH regression estimator as:

$$\boxed{\hat{m}_{jk} = \frac{\sum_a \sum_b w_{ab} \nu_{j+a,k+b} \bar{z}_{j+a,k+b}}{\sum_a \sum_b w_{ab} \nu_{j+a,k+b}} .}$$

2.4 ASH Regression Extensions

REMARK 1: For the survey sampled data, each data point takes the extended form

$$\{(x, y, z, \alpha)_i, \quad i = 1, \ldots, n\} ,$$

where α_i is the effective sampling weight. Previously, we have assumed that $\alpha_i = 1$ for all cases. Here, the frequency counts ν_{jkl} are replaced by the sum of these α_i weights rather than 1's.

REMARK 2: Occasionally, our data will include other covariates and be of the form

$$\{(x, y, z, t, \alpha)_i, \quad i = 1, \ldots, n\} ,$$

where t is some covariate of interest. Then we compute the ASH regression estimator $\hat{m}(x, y, t)$ by simply adding another loop to the numerator and denominator of the $\hat{m}_{jk}$ equation above. The sampling weights are the same of course. What could be easier? Typically, we will map the estimate at several levels of t, for example, $\hat{m}(x, y, t = t_0)$.

REMARK 3: The 1-dimensional ASH regression prescription was first published in Härdle and Scott (1992) under the name WARPing.

3 Mapping Details

After the "usa()" is plotted, the regression ASH is computed over the entire 300×200 mesh and added to the figure by using either the Splus "contour" or "image" function and the argument "add=T". Typically, the contour lines will extend slightly outside the US borders. A simple trick removes those lines, by applying "polygon" to two pieces that outline half the borders of the US and the surrounding rectangles. This will be illustrated in the examples.

The internal boundary solution is not handled in an elegant fashion currently. Thresholding could be applied, but we find the problem is relatively localized and have left it for the reader to discover. A bootstrap algorithm has been implemented to estimate the pointwise error. We have used this to replace or delete regions where the estimator behaves erratically.

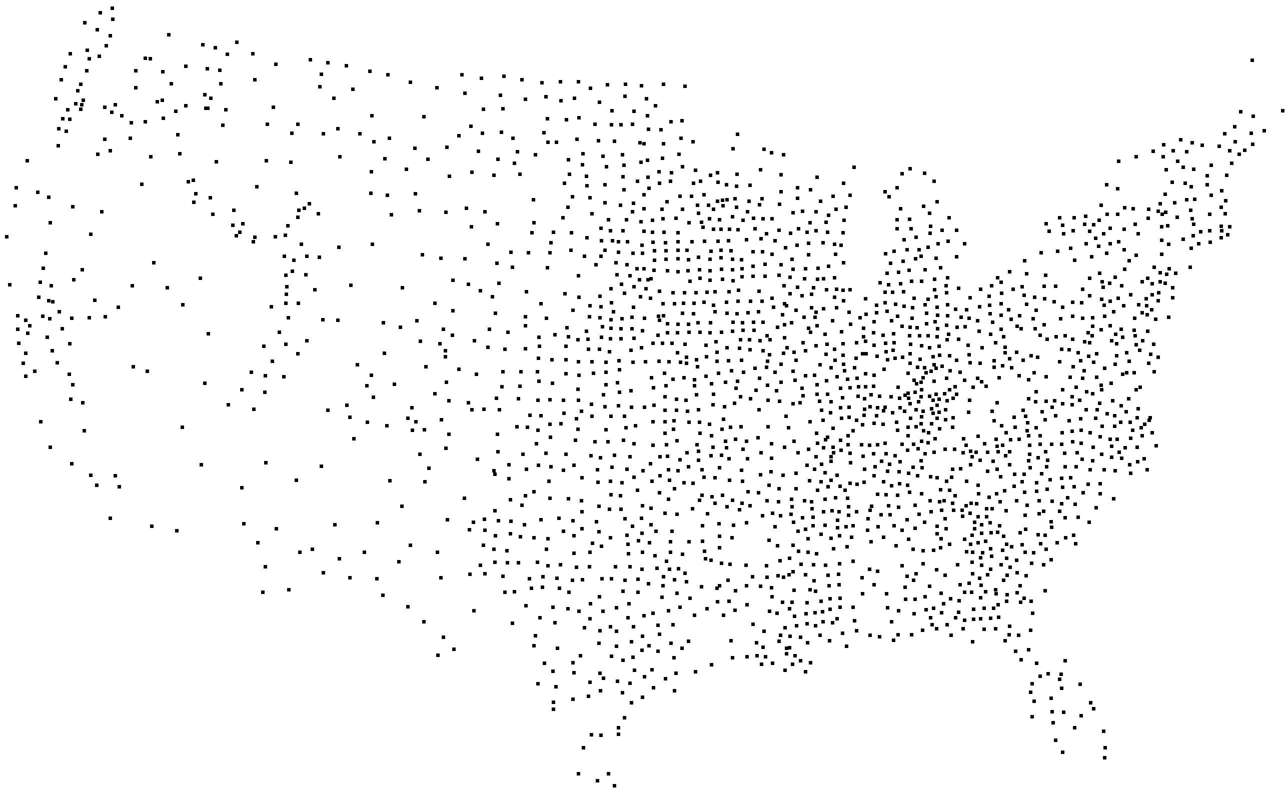

Figure 1: Centers of US Counties for Sampling

4 Example

The data considered in this section come from the Farm Costs and Returns Survey. This is a stratified complex design survey which is used to measure finances and production of all U.S. agriculture. The weight of each observation was taken to be the inverse of the probability of selection.

A sample of $n = 13{,}000$ of 1.7 million farms was drawn. For these data, the FIPS code for each observation was known. Thus the exact location of each observation was assigned to the location of the population centroid of the county where the farm is located. The map of the 3,100 centroids is shown in Figure 1. Observe that the resolution is much greater east of the Mississippi.

When loess (Cleveland, 1979), kriging, and other methods were applied to these data, each failed to produce a usable surface from the data. The result was always a smooth surface for most of the country with an enormous peak at an edge. However, the ASH regression algorithm with $m_x = m_y = 5$ produced excellent results.

Figure 2 displays our ASH estimate of the fraction of government payments to gross farm income. For the most part, the value of the regression surface is quite small near the US borders, except in Texas and along a portion of the border with Canada (where government subsidies are even greater!). We do not find the bias incurred particularly misleading. The patterns in this figure are quite interesting to policy makers.

There are a number of other issues that deserve exploration, such as the practical effect of the sampling weights, the existence of internal boundaries, and the addition of covariates. These are

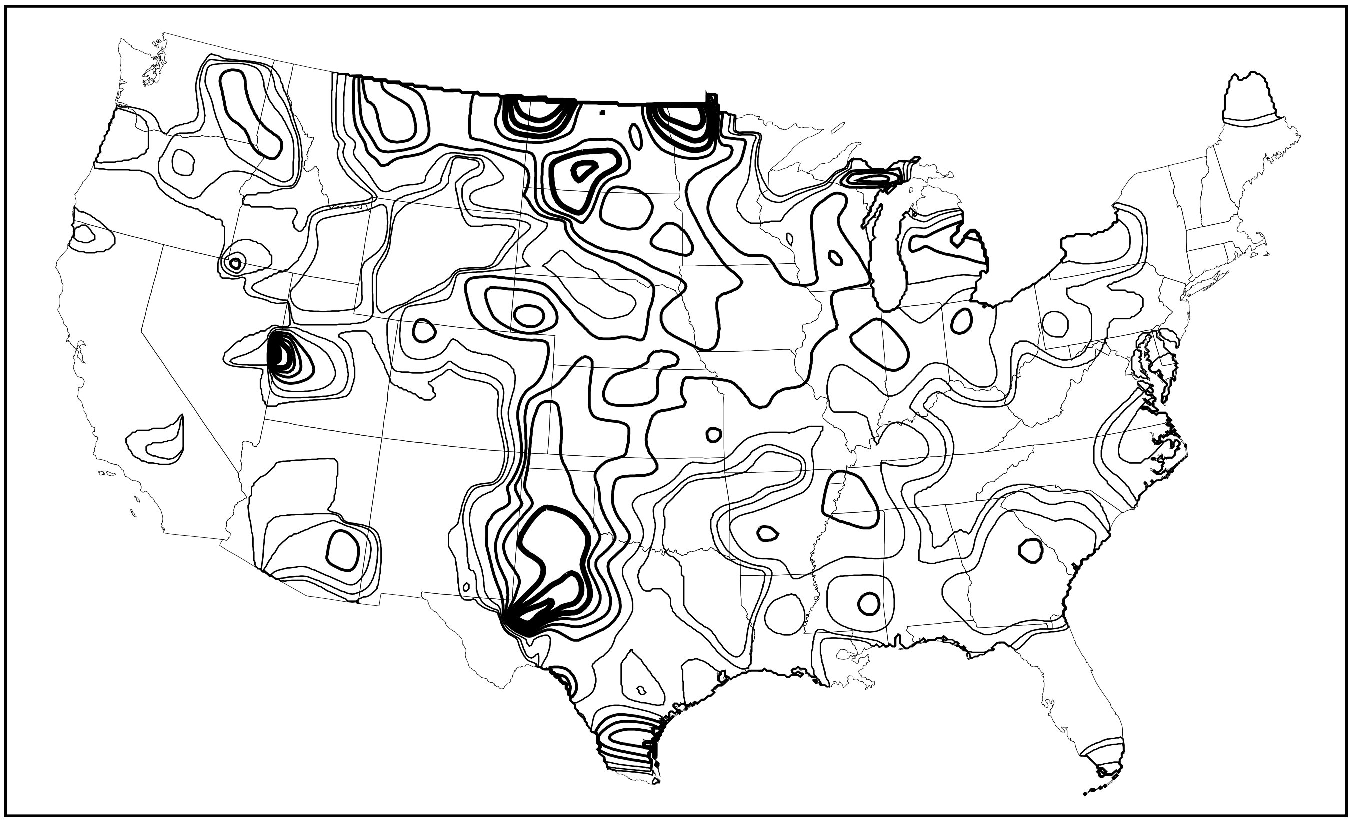

Figure 2: Contours of the farm subsidy ASH estimates: 9 levels

explored in a companion paper (Whittaker and Scott, 1992).

5 Discussion

The naive ASH is not robust, but is easily adapted to handle weighted data and covariates with small computational overhead. Elegant procedures without covariate handling have been considered by Tobler (1979). We have not taken advantage of possible small gains available by considering spatial correlations.

However, kriging and lowess both produced estimates with huge values at the boundary and outside the US borders. Apparently, the trick of placing a rectangular grid on the US extending outside the borders fails because the algorithms require explicit knowledge of the boundary locations as input.

The actual proximate reason for failure, interestingly enough, is due to the "adaptive" nature of these algorithms, which fit the LPR over a region with a certain fraction of the data. In places where the mesh extends offshore, the regression estimate is reaching far inland for any data to fit — the extrapolation problem once again. (Explicit boundary handling would fix this, presumably).

The ASH procedure used a fixed (or nonadaptive) neighborhood. The result is regions where the regression estimate is undefined (0/0). However, we are more comfortable with such undefined regions than with providing dubious estimates obtained by spanning empty spaces.

6 References

Carr, D.B. (1990). "Looking at Large Data Sets Using Binned Data Plots," PNL-7301, Battelle Labs, Richland, WA.

Cleveland, W.S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association*, 87:829–836.

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

Fan, J. (1992). "Design Adaptive Nonparametric Regression." *Journal of the American Statistical Association*, 87:998–1004.

Fan, J. and Marron, J.S. (1994). "Fast Implementations of Nonparametric Curve Estimators." *Journal of Computational and Graphical Statistics*, 3:35–56.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Boston.

Härdle, W. and Scott, D.W. (1992). "Smoothing by Weighted Averaging of Shifted Points," *Computational Statistics*, 7:97–128.

Hastie, T. and Tibsharani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Nadaraya, E.A. (1964). "On Estimating Regression," *Theory of Probability and its Applications*, 9:141–142.

Watson, G.S. (1964). "Smooth Regression Analysis," *Sankhy A*, 26:359–372.

Scott, D.W. (1983). "Nonparametric Probability Density Estimation for Data Analysis in Several Dimensions," *Proceedings of the Twenty-Eighth Conference on the Design of Experiments in Army Research Development and Testing*, pp. 387–397.

Scott, D.W. (1985). "Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," *Annals of Statististics*, 13:1024–1040.

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

Tobler, W.R. (1979). "Smooth Pycnophylactic Interpolation for Geographical Regions," *Journal of the American Statistical Association*, 74:519–530.

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: 8-17.