Nonparametric Small Area Estimation Using Penalized Spline Regression

J. D. Opsomer
Iowa State University*

G. Claeskens
Katholieke Universiteit Leuven

M. G. Ranalli
Colorado State University

G. Kauermann
Universität Bielefeld

F. J. Breidt
Colorado State University

10th January 2005

Abstract

We propose a new small area estimation approach that combines small area random effects with a smooth, nonparametrically specified trend. By using penalized splines as the representation for the nonparametric trend, it is possible to express the small area estimation problem as a mixed effect model regression. This model is readily fitted using existing model fitting approaches such as restricted maximum likelihood. We develop a corresponding bootstrap approach for model inference and estimation of the small area prediction mean squared error. The applicability of the method is demonstrated on a survey of lakes in the Northeastern US.

Key Words: mixed model, best linear unbiased prediction; bootstrap inference, natural resource survey.

1 Introduction

In many surveys, it is of interest to provide estimates for small domains within the overall population of interest. Depending on the overall survey sample size, design-

*Department of Statistics, Iowa State University, Ames, IA 50014, USA; jopsomer@iastate.edu.

based inference methods might not be appropriate for all or some of these small domains, so that survey practitioners have often resorted to model-based estimators in this case. The term "small area estimation" is often used to denote this kind of estimation setting. Ghosh and Rao (1994) give an overview of the most commonly used types of estimators used by survey statisticians, including synthetic and composite estimators, mixed model prediction, and Bayesian approaches. To date, the approaches in use by survey statisticians have relied on parametric, most often linear, modelling techniques. In this article, we propose a new type of small area estimator that relies on a nonparametric model formulation.

In principle, a nonparametric model might have significant advantages compared to current fully parametric approaches when the functional form of the relationship between the variable of interest and the covariates cannot be specified a priori, since erroneous specification of the model can result in biased estimators. Even when a specific functional form appears reasonable, the nonparametric model provides a more robust model alternative that can be useful in the process of model checking and validation. Despite these possible advantages, nonparametric approaches have not made inroads in small area estimation, due in large part to the methodological difficulties of incorporating existing smoothing techniques into the estimation tools used by survey statisticians.

Penalized spline regression, often referred to as *P-splines*, is a nonparametric method recently popularized by Eilers and Marx (1996). P-splines are an attractive smoothing method, because of their flexibility and the ability to incorporate them into a large range of modelling contexts. We refer to Ruppert et al. (2003) for an overview of applications of P-splines to different settings. Because of their close connections with linear mixed models discussed in Wand (2003), P-splines are also a natural candidate for constructing nonparametric small area estimators, as we show in the current article. In doing so, we extend the mixed model small area estimation approach described in Battese et al. (1988) to the setting in which the mean function can be nonparametrically (or semiparametrically) specified.

The ability to combine nonparametric regression and mixed model regression with P-splines has recently been used in other contexts. Parise et al. (2001), Coull et al. (2001), Coull et al. (001a) and Liang (2003) all provide examples of using penalized splines in the construction of mixed effect regression models for the analysis of data containing random effects. In the survey context, Zheng and Little (2003) propose a model-based estimator for cluster sampling, in which the regression model combines a spline model with a random effect for the clusters. Our approach is

conceptually similar to that of these other authors, but targeted specifically to small area estimation. We show consistency of the estimator, compute its mean squared error and provide tests for small area effects and non-linearities, both via asymptotic theory and bootstrap methods.

We illustrate the applicability of the nonparametric small area estimation approach on a survey of lakes in the Northeastern states of the U.S. In that survey, 334 lakes were sampled from a population of 21,026 lakes. We use small area estimation to produce estimates of mean *acid neutralizing capacity* (ANC) for each of 113 8-digit *Hydrologic Unit Codes* (HUC) in the region. In this application, we show how the inclusion of a spatial spline can improve the fit relative to a model which only uses a random effect for the small areas, as would be done in traditional small area estimation. We also argue that the model that includes both the spatial spline and a HUC effect performs better than a model that only includes a spline, at least in the small area estimation context.

Section 2 describes the proposed nonparametric small area estimation method. The theoretical properties and inference results are in Section 3. Finally, the application is presented in Section 4.

2 Description of Methodology

We begin by describing the spline-based nonparametric regression model and estimator outside of the small area context. We closely follow the description in Ruppert et al. (2003). Consider first the simple model

$$y_i = m_o(x_i) + \varepsilon_i,$$

where the ε_i are independent random variables with mean zero and variance σ^2_ε. The function $m_o(\cdot)$ is unknown, but if this function is to be estimated using P-splines, we assume that it can be approximated sufficiently well by

$$m(x; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - \kappa_k)^p_+. \tag{1}$$

Here p is the degree of the spline, $(x)^p_+$ denotes the function $x^p \mathrm{I}_{\{x>0\}}$, $\kappa_1 < \ldots < \kappa_K$ is a set of fixed *knots* and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)', \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$ are the coefficient vectors for the "parametric" and the "spline" portions of the model, respectively. If K is sufficiently large (guidelines are given below), the class of functions $m(x; \boldsymbol{\beta}, \boldsymbol{\gamma})$

is very large and can approximate most smooth functions $m_o(\cdot)$ with a very high degree of accuracy, even for p small (say, between 1 and 3). As is commonly done in the P-spline context, we assume that the lack-of-fit error $m_o(\cdot) - m(\cdot;\boldsymbol{\beta},\boldsymbol{\gamma})$ is negligible relative to the estimation error $m(\cdot;\boldsymbol{\beta},\boldsymbol{\gamma}) - m(\cdot;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$.

The spline function (1) uses the *truncated polynomial spline basis* $\{1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p\}$ to approximate the function m_0. Other bases are also possible and, especially when x is multivariate, might be preferable to the truncated polynomials. Regardless of the choice of basis, the spline function can be expressed as a linear combination of basis functions. In Section 4, we introduce the radial basis functions for use in the spatial context.

In P-spline regression, K is typically taken to be large relative to the size of the dataset, with 1 knot every 4 or 5 observations, say. Hence, the model (1) is potentially over-parameterized. This issue is avoided by putting a *penalty* on the magnitude of the spline parameters $\boldsymbol{\gamma}$. For a given dataset $\{(x_i, y_i) : i = 1, \ldots, n\}$, this is done by defining the regression estimators as the minimizers over $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of

$$\sum_{i=1}^{n} \left(y_i - m(x_i;\boldsymbol{\beta},\boldsymbol{\gamma})\right)^2 + \lambda_\gamma \boldsymbol{\gamma}'\boldsymbol{\gamma}.$$

where λ_γ is a fixed penalty parameter. However, different values of λ_γ result in different estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, so that it is of interest to treat λ_γ as an unknown parameter as well. As discussed in Ruppert et al. (2003), this can be conveniently done by treating the $\boldsymbol{\gamma}$ as a random effect in a linear mixed model specification, which will allow joint estimation of λ_γ, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by maximum likelihood methods.

In small area estimation, a commonly used approach is to express the relationship between the variable of interest and any auxiliary variables as a linear model supplemented by a random effect for the small areas (e.g. Battese et al. 1988). Since both the P-spline and the small area estimation models can be viewed as random effects models, it is natural to try to combine both into a nonparametric small area estimation framework based on linear mixed model regression.

Specifically, suppose there are T small areas, $U_1, \ldots, U_T$, for which estimates are to be constructed. Define $d_{it} = \mathrm{I}_{\{i \in U_t\}}$, and for each observation i, let $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{iT})$. Let $\boldsymbol{Y} = (y_1, \ldots, y_n)'$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & & & \vdots \\ 1 & x_n & \cdots & x_n^p \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} (x_1-\kappa_1)_+^p & \cdots & (x_1-\kappa_K)_+^p \\ \vdots & & \vdots \\ (x_n-\kappa_1)_+^p & \cdots & (x_n-\kappa_K)_+^p \end{bmatrix}$$

and $\boldsymbol{D} = (\boldsymbol{d}_1', \ldots, \boldsymbol{d}_n')'$. If other variables are available that need to be included in the model as parametric terms, they can be added into the $\boldsymbol{X}$ fixed effect matrix. We assume that the data follow the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{D}\boldsymbol{u} + \boldsymbol{\varepsilon} \tag{2}$$

where

$$\begin{aligned} \boldsymbol{\gamma} &\sim (\boldsymbol{0}, \boldsymbol{\Sigma}_\gamma) \text{ with } \boldsymbol{\Sigma}_\gamma \equiv \sigma_\gamma^2 \boldsymbol{I}_K \\ \boldsymbol{u} &\sim (\boldsymbol{0}, \boldsymbol{\Sigma}_u) \text{ with } \boldsymbol{\Sigma}_u \equiv \sigma_u^2 \boldsymbol{I}_T \\ \boldsymbol{\varepsilon} &\sim (\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon) \text{ with } \boldsymbol{\Sigma}_\varepsilon \equiv \sigma_\varepsilon^2 \boldsymbol{I}_n \end{aligned} \tag{3}$$

and each of the random components is assumed independent of the others.
The model (2) includes both the spline function, which can be thought of as a non-parametric mean function specification and includes $\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}$, and the small area random effects $\boldsymbol{D}\boldsymbol{u}$. For the purpose of fitting this model and using the appropriate amount of smoothing for the spline, it is convenient to continue to treat $\boldsymbol{Z}\boldsymbol{\gamma}$ as a random effect term, so that $\text{Var}(\boldsymbol{Y}) \equiv \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\Sigma}_\gamma\boldsymbol{Z}' + \boldsymbol{D}\boldsymbol{\Sigma}_u\boldsymbol{D}' + \boldsymbol{\Sigma}_\varepsilon$.
If the variances of the random components are known, standard results from BLUP theory (e.g. McCulloch and Searle, 2001, Chapter 9) guarantee that, given the model specifications (2) and (3), the GLS estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} \tag{4}$$

and the predictors

$$\begin{aligned} \widehat{\boldsymbol{\gamma}} &= \boldsymbol{\Sigma}_\gamma\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ \widehat{\boldsymbol{u}} &= \boldsymbol{\Sigma}_u\boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \end{aligned} \tag{5}$$

are optimal among all linear estimators/predictors.
For a given small area U_t, we assume that we are interested in predicting

$$\bar{y}_t = \bar{\boldsymbol{x}}_t\boldsymbol{\beta} + \bar{\boldsymbol{z}}_t\boldsymbol{\gamma} + u_t, \tag{6}$$

where $\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{z}}_t$ are the true means of the powers of x_i (up to p) and of the spline basis functions over the small area, and u_t is the small area effect. Both $\bar{\boldsymbol{x}}_t$ and $\bar{\boldsymbol{z}}_t$ are assumed known. Note that $\bar{y}_t$ is not generally equal to the true mean of the y_i in small area U_t, because it ignores the mean of the errors ε_t. The difference between both quantities is usually ignored in practice, and we will do the same here.

Clearly, $u_t = \bar{\boldsymbol{d}}_t \boldsymbol{u} = \boldsymbol{e}_t \boldsymbol{u}$, where $\boldsymbol{e}_t$ is a vector with 1 in the tth position and 0s everywhere else. As a predictor of $\bar{y}_t$, we therefore use

$$\widehat{y}_t = \bar{\boldsymbol{x}}_t \widehat{\boldsymbol{\beta}} + \bar{\boldsymbol{z}}_t \widehat{\boldsymbol{\gamma}} + \boldsymbol{e}_t \widehat{\boldsymbol{u}}, \tag{7}$$

which is a linear combination of the GLS estimator (4) and the BLUPs in (5), so that $\widehat{y}_t$ is itself BLUP for $\bar{y}_t$.

If the variances are unknown, EBLUP versions of (4), (5) and (7) are constructed by replacing $\sigma^2_\gamma, \sigma^2_u, \sigma^2_\varepsilon$ by estimators. Estimated parameters (4) and predictions (5) can be obtained by *Restricted Maximum Likelihood* (REML) minimization or related methods (Patterson and Thompson, 1971), which are implemented in `PROC MIXED` in `SAS` and `lme()` in `S-Plus` and `R`, among others.

3 Theoretical Properties

3.1 Prediction Mean Squared Error

We consider the prediction error $\widehat{y}_t - \bar{y}_t$ in the case of known variance components. To simplify the expressions, we let $\boldsymbol{W} = [\boldsymbol{Z}\boldsymbol{D}]$, $\boldsymbol{\omega} = (\boldsymbol{\gamma}', \boldsymbol{u}')'$, $\bar{\boldsymbol{w}}_t = (\bar{\boldsymbol{z}}_t, \boldsymbol{e}_t)$ and

$$\boldsymbol{\Sigma}_w = \begin{bmatrix} \boldsymbol{\Sigma}_\gamma & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_u \end{bmatrix}.$$

Then,

$$\widehat{y}_t - \bar{y}_t = \boldsymbol{c}_t \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) + \bar{\boldsymbol{w}}_t \left(\boldsymbol{\Sigma}_w \boldsymbol{W}' \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{\omega} \right) \tag{8}$$

with $\boldsymbol{c}_t = \bar{\boldsymbol{x}}_t - \bar{\boldsymbol{w}}_t \boldsymbol{\Sigma}_w \boldsymbol{W}' \boldsymbol{V}^{-1} \boldsymbol{X}$. This expression can be used to derive the properties of the small area predictors under different frameworks.

If both the spline coefficients and the small areas are treated as true random effects in the underlying model (2), the mean prediction error is 0 and the covariance between the two terms in (8) is also 0, so that mean squared error (MSE) of the prediction errors is readily calculated to be

$$\mathrm{E}(\widehat{y}_t - \bar{y}_t)^2 = \boldsymbol{c}_t (\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{c}_t' + \bar{\boldsymbol{w}}_t \boldsymbol{\Sigma}_w \left(\boldsymbol{I} - \boldsymbol{W}' \boldsymbol{V}^{-1} \boldsymbol{W} \boldsymbol{\Sigma}_w \right) \bar{\boldsymbol{w}}_t'. \tag{9}$$

This expression corresponds to equation (3.6) in Battese et al. (1988).

If the variances of the random effects are estimated from the data and an EBLUP is constructed, expression (9) is no longer equal to the MSE of the prediction errors. Using the results of Jiang (1998) on consistency of the REML parameter estimators,

it can be shown that, as long as the distribution functions of the errors and the random effects are symmetric and the derivatives of $\boldsymbol{\beta}$ and $\boldsymbol{V}^{-1}$ with respect to the variance parameters exist and are bounded, (9) is the variance of the asymptotic distribution of $\widehat{y}_t - \bar{y}_t$. Under those same conditions, this asymptotic prediction MSE is consistently estimated by replacing all unknown quantities in (9) by their REML estimators (Patterson and Thompson, 1971).

3.2 Testing for small area effects and non-linearities

In the model, there are two main sources of variability (not counting the error terms): (i) the small area effects, and (ii) the deviation from the parametric pth degree polynomial model, as accounted for by the spline functions. Since both of these features are modeled via random effects in a mixed linear model, the absence of one of the effects is characterized by the zero-ness of the corresponding variance component. A likelihood ratio test (or restricted likelihood ratio test) for testing the presence of small area effects is readily constructed. To test the hypothesis $H_{0,u} : \sigma_u^2 = 0$ versus the one-sided alternative $H_{a,u} : \sigma_u^2 > 0$ we fit the model twice, once without the small area random effects, resulting in the likelihood (or restricted likelihood) value $\mathcal{L}_0$, and once with these random effects included, giving $\mathcal{L}_1$. The test statistic equals $\mathcal{L}_u = 2\{\mathcal{L}_1 - \mathcal{L}_0\}$. Similarly, a (restricted) likelihood ratio statistic to test for the presence of any structure more complicated than a pth degree polynomial, $H_{0,\gamma} : \sigma_\gamma^2 = 0$ versus $H_{a,\gamma} : \sigma_\gamma^2 > 0$ is denoted by $\mathcal{L}_\gamma$. It is also possible to test for both effects simultaneously, more precisely, $H_0 : \sigma_u^2 = 0, \sigma_\gamma^2 = 0$ versus $H_a : \sigma_u^2 > 0$ or $\sigma_\gamma^2 > 0$.

Define $\boldsymbol{\lambda} = (\lambda_\gamma, \lambda_u)$, $\boldsymbol{\lambda}_0$ its value under the null hypothesis for any of the three hypotheses, the rescaled variance matrix $\boldsymbol{V}_\lambda = \boldsymbol{V}/\sigma_\varepsilon^2$ and the projection matrix $\boldsymbol{Q}(\boldsymbol{\lambda}) = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{t}}\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}_\lambda^{-1}$. Denote $\boldsymbol{Z}_1 = \boldsymbol{Z}$ and $\boldsymbol{Z}_2 = \boldsymbol{D}$. Define further the 2×2 matrix $\boldsymbol{G}_n$ with entries $G_{n,k\ell} = \mathrm{tr}\{\boldsymbol{Z}_k\boldsymbol{Z}_k^t\boldsymbol{V}_{\lambda_0}^{-1}\boldsymbol{Q}(\boldsymbol{\lambda}_0)\boldsymbol{Z}_\ell\boldsymbol{Z}_k^t\boldsymbol{V}_{\lambda_0}^{-1}\boldsymbol{Q}(\boldsymbol{\lambda}_0)\}$.

Theorem 3.1 *Assume that the number of small areas $T = T_n \to \infty$, and the number of knots $K = K_n \to \infty$ such that $T_n = o(n)$ and $K_n = o(n)$. Assume, too, that for $j = 1, 2$ the limit $tr\{(\boldsymbol{Z}_j^t\boldsymbol{V}_{\lambda_0}^{-1}\boldsymbol{Q}(\boldsymbol{\lambda}_0)\boldsymbol{Z}_j)^2\} \to \infty$ holds, and that*

$$tr\{(\boldsymbol{Z}_j^t\boldsymbol{V}_{\lambda_0}^{-1}\boldsymbol{Q}(\boldsymbol{\lambda}_0)\boldsymbol{Z}_j)^4\}/tr\{(\boldsymbol{Z}_j^t\boldsymbol{V}_{\lambda_0}^{-1}\boldsymbol{Q}(\boldsymbol{\lambda}_0)\boldsymbol{Z}_j)^2\}^2 \to 0.$$

Then, with $\boldsymbol{\lambda}_0 = (\lambda_{\gamma,0}, 0)$ to test $H_{0,u}$ (resp. $\boldsymbol{\lambda}_0 = (0, \lambda_{u,0})$ to test $H_{0,\gamma}$), the (restricted) likelihood ratio statistic $\mathcal{L}_u$ (resp. $\mathcal{L}_\gamma$) has an asymptotic distribution which

is an equal mixture of a point mass at zero and a chi-squared with one degree of freedom, denoted $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.
To test H_0 *that both variance components are zero,* $\boldsymbol{\lambda}_0 = (0, 0)$ *and the (restricted) likelihood ratio statistic has asymptotically the mixture distribution* $(0, N_1^2, (N_1 - sN_2)^2/(1 + s^2), N_1^2 + N_2^2)$ *with probabilities* $(1/2 - r, 1/4, 1/4, r)$ *where* $(N_1, N_2) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I}_2)$, *and* $s = \lim_{n\to\infty} s_n$, $r = \lim_{n\to\infty} r_n$ *with* $s_n = G_{n,12}/\sqrt{|\boldsymbol{G}_n|}$ *and* $r_n = \cos^{-1}\left(s_n/\sqrt{1 + s_n^2}\right)/(2\pi)$.

Proof. To prove the first part of the theorem, we follow the same line of arguments as to prove Theorem 2 of Claeskens (2004), with the difference that only one variance component is set to zero. The simplification $\boldsymbol{V}_{\lambda_0} = \boldsymbol{I}$ only occurs in the proof of the last part, which follows immediately from that Theorem 2. As in standard testing problems without boundary parameters (see Ferguson, 1996, Chapter 22) the asymptotic distribution is the same as if there were no nuisance parameters. ■

The first assumption on the trace guarantees that the Fisher information matrix is positive definite for large n, and could be weakened to requiring just this. The second assumption on the trace implies that the standardized score converges in distribution to a standard normal random variable, which leads to the χ_1^2 component in the mixture distribution. This condition holds if the eigenvalues of $\boldsymbol{Z}_j^t \boldsymbol{V}_{\lambda_0}^{-1} \boldsymbol{Q}(\boldsymbol{\lambda}_0) \boldsymbol{Z}_j$ are $O(n^\zeta)$ with $\zeta \geq 0$ and the Fisher information matrix is positive definite.

3.3 Bootstrapping small area and local effects

As discussed in Crainiceanu and Ruppert (2004) in the spline regression context, the finite sample distribution of the likelihood ratio statistic is often poorly approximated by these asymptotic distribution, so that it is useful to supplement it by a bootstrap-based procedure. Bootstrap replicate observations are generated as

$$\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\boldsymbol{\gamma}^* + \boldsymbol{D}\boldsymbol{u}^* + \boldsymbol{\varepsilon}^*, \tag{10}$$

where $\boldsymbol{\gamma}^*$, $\boldsymbol{u}^*$ and $\boldsymbol{\varepsilon}^*$ are bootstrap replicates of the random components in the model. In principle there are various possibilities to draw such replicates. A natural way to do this is to make use of the stochastic model given in (3) with fitted variance parameters. This requires that specific parametric distributions for the random components be chosen, and in practice, Gaussian distributions would often be used for this purpose. While straightforward to implement, this approach could lead to biased inference if these parametric distributions are misspecified. A more robust

alternative which is pursued here is to resample from the empirical distributions of the fitted residuals and predictions, suitably adjusted so that their first two moments match the desired ones.

We start from the predictors $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{u}}$ obtained in (5). Note that the variance of $\widehat{\boldsymbol{\gamma}}$ is

$$\begin{aligned} \mathrm{Var}(\widehat{\boldsymbol{\gamma}}) &= \boldsymbol{\Sigma}_\gamma \boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{V}(\boldsymbol{I}-\boldsymbol{Q})'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_\gamma \\ &= \sigma_\gamma^4 \boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{Z} \end{aligned}$$

so that we need to adjust $\widehat{\boldsymbol{\gamma}}$ to obtain predictors with the correct second moment by letting

$$\tilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}\ (\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{Z})^{-1/2}/\sigma_\gamma.$$

As an alternative, it might be possible to use $\tilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}(\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1/2}/\sigma_\gamma$, if the error in estimating $\boldsymbol{\beta}$ is negligible relative to that in predicting $\boldsymbol{\gamma}$. The same reasoning leads to

$$\tilde{\boldsymbol{u}} = \widehat{\boldsymbol{u}}(\boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{D})^{-1/2}/\sigma_u$$

(possibly removing $\boldsymbol{Q}$ again). Finally, to generate estimated errors, we start from

$$\begin{aligned} \widehat{\boldsymbol{\varepsilon}} &= \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{\gamma}} - \boldsymbol{D}\widehat{\boldsymbol{u}} \\ &= \boldsymbol{\Sigma}_\varepsilon \boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{Y} \end{aligned}$$

so that

$$\mathrm{Var}(\widehat{\boldsymbol{\varepsilon}}) = \sigma_\gamma^4 \boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{Q}).$$

Since we only need to have the correct second moments element-wise, it is sufficient to adjust the estimated errors as follows:

$$\tilde{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\varepsilon}}\ \mathrm{diag}\{\boldsymbol{V}^{-1}(\boldsymbol{I}-\boldsymbol{Q})\}^{-1/2}/\sigma_\varepsilon.$$

We can again choose to remove $\boldsymbol{Q}$. Bootstrap resampling is done from $\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{\varepsilon}}$ after centering to obtain zero-mean random components.

The Mean Squared Error (9) can now be bootstrapped by taking the bootstrap mean

$$E^*(\widehat{y}_t^* - \bar{y}_t^*)^2 \tag{11}$$

where $\bar{y}_t^* = \bar{x}_t\widehat{\beta} + \bar{z}_t\gamma^* + u_t^*$ and $\widehat{y}_t^*$ is the predicted value based on the bootstrap observations $\boldsymbol{Y}^*$. The superscript $*$ in (11) indicates that the mean is taken with

respect to the bootstrap distribution. In practice, this expectation step is usually replaced by taking the average over a number of bootstrap replicates, that is

$$\frac{1}{B}\sum_{b=1}^{B}(\widehat{y}_t^{*b} - \bar{y}_t^{*b})^2$$

where B is the total number of bootstrap replicates and superscript $*b$ refers to the b-th bootstrap simulation. If desired, the computational effort of the bootstrap procedure could be reduced by keeping the estimates of the variance parameters fixed across the bootstrap replicates, instead of fitting the full REML procedure for each replicate.

We illustrate the use of the bootstrap procedure in approximating the distribution of the likelihood ratio statistic for the case of testing $H_{0,u} : \sigma_u^2 = 0$. First, we fit the model with $H_{0,u} : \sigma_u^2 = 0$ and the alternative model $H_{1,u} : \sigma_u^2 \geq 0$ to the data and obtain the likelihood (or restricted likelihood) statistic $\mathcal{L}_u = 2\{\mathcal{L}_1 - \mathcal{L}_0\}$. To assess the significance of $\mathcal{L}_u$, the distribution of $\mathcal{L}_u$ under $H_{0,u}$ is approximated by generating bootstrap replicates as

$$\boldsymbol{Y}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon}^*, \tag{12}$$

where $\boldsymbol{\gamma}^*$ and $\boldsymbol{\varepsilon}^*$ are generated as discussed above. For each bootstrap replicate sample we fit both models, including the estimation of the variance components. Denoting with $\mathcal{L}_u^{*b}$ the value of $\mathcal{L}_u$ in the b-th bootstrap, $b = 1, \ldots, B$, one can assess the significance of $\mathcal{L}_u$ using the empirical distribution of $\mathcal{L}_u^{*b}$.

4 Application

Between 1991 and 1996, the Environmental Monitoring and Assessment Program (EMAP) of the U.S. Environmental Protection Agency conducted a survey of lakes in the Northeastern states of the U.S. The survey is based on a population of 21,026 lakes from which 334 lakes were surveyed, some of which were visited several times during the study period. The total number of measurements is 551. Figure 1 shows the region of interest and the locations of the sampled lakes. We refer to Messer et al. (1991) and Larsen et al. (2001) for a description of the EMAP program and the Northeastern Lakes survey.

In this article, we consider the estimation of the mean *acid neutralizing capacity* (ANC) for each of 113 small areas defined by 8-digit Hydrologic Unit Codes (HUC)

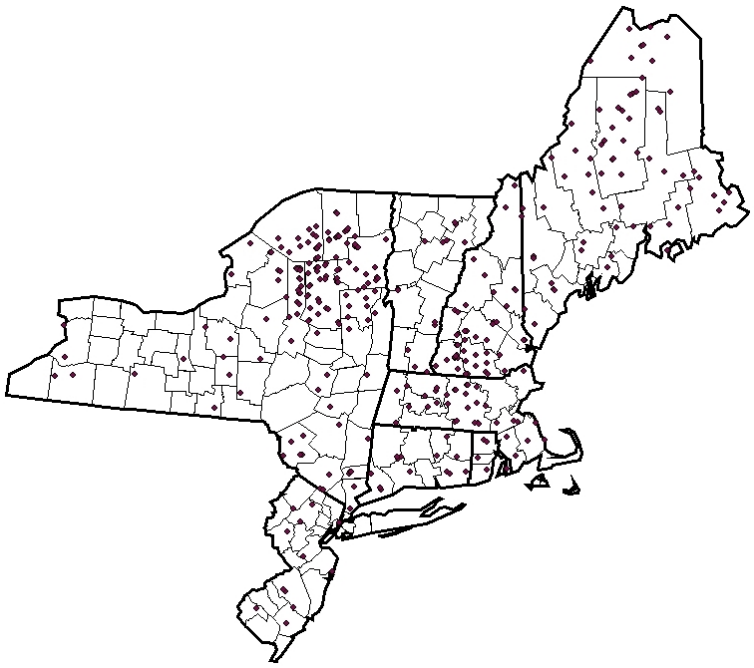

Figure 1: Locations of sampled lakes in Northeastern U.S.

within the region of interest. ANC, also called *acid binding capacity* or *total alkalinity*, measures the buffering capacity of water against negative changes in pH (Wetzel, 1975, p. 172), and is often used as an indicator of the acidification risk of water bodies in water resource surveys. Figure 2 displays a map of the HUCs in the region of interest, with the average ANC computed for all HUCs in which sample observations were located. The map also shows the locations of the 27 HUCs in which no sample observations are available.

The variables that can be used in the construction of a small area estimation model in this application are the geographical coordinates of the centroid of each lake (in the UTM coordinate system) and its elevation. After trying different combinations of parametric and nonparametric specifications for these variables, it was determined that a bivariate spline on the UTM coordinates and a linear term for elevation provided the best model fit. We will therefore describe the construction of the small area estimator for this combination of terms.

In principle, the spline function (1) could be extended to the bivariate case by taking tensor products of basis functions in the North/South and East/West directions. However, this leads to very large numbers of basis functions and numerical instability in the fitting algorithm. Instead, we will follow Ruppert et al. (2003, p.253) in using

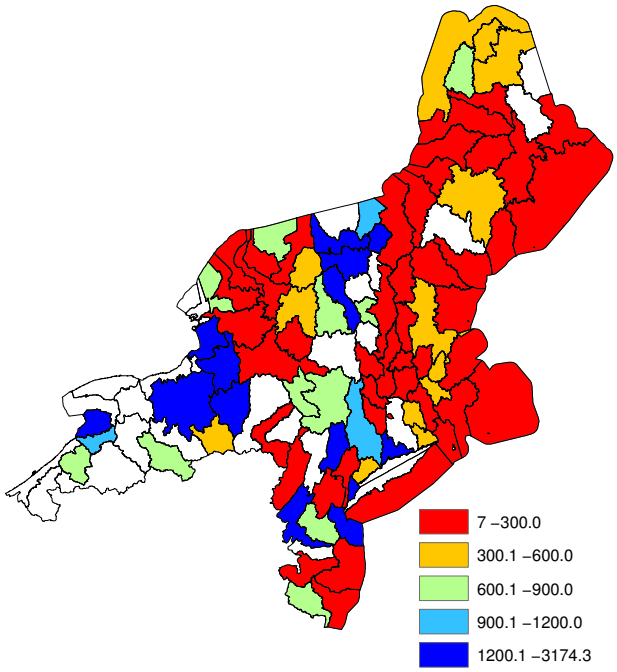

Figure 2: Hydrologic Unit Code (HUC) small areas within Northeastern U.S. region, with average ANC computed in all small areas containing sample observations.

a *transformed radial basis*, defined as

$$\boldsymbol{Z} = [C(\boldsymbol{x}_i - \boldsymbol{\kappa}_k)]_{\substack{1 \le i \le n \\ 1 \le k \le K}} [C(\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'})]^{-1/2}_{1 \le k,k' \le K}, \tag{13}$$

where $C(\boldsymbol{r}) = ||\boldsymbol{r}||^2 \log ||\boldsymbol{r}||$, $\boldsymbol{x}_i = (x_{1i}, x_{2i})$ denotes the geographical coordinates for observation i and $\boldsymbol{\kappa}_k, k = 1, \ldots, K$ are spline knots. The multiplication by $[C(\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'})]^{-1/2}$ is necessary in order to allow the coefficients of the basis functions to be specified in the model as being independent and identically distributed random effects. The locations of the 80 knots are selected by the space-filling algorithm implemented in the `cover.design()` function in the `FUNFITS` package for `S-plus` (Nychka et al. 1998). Figure 3 shows the locations of the knots selected by this approach.

The ANC small area model can now be written as in (2) with variance components (3). That model includes $\boldsymbol{Y}$ for the ANC observations, $\boldsymbol{X}$ a matrix containing an intercept and the linear elevation term, $\boldsymbol{Z}$ as in (13) for the spatial locations, and $\boldsymbol{D}$ a matrix of indicators for the HUCs. The model is fitted using REML as implemented in `lme()` in `S-plus`. The parameter estimates and corresponding P-values are shown in Table 1. The P-values were computed using the resampling bootstrap procedure described in Section 3.3 and a bootstrap sample size $B = 1000$. We implemented

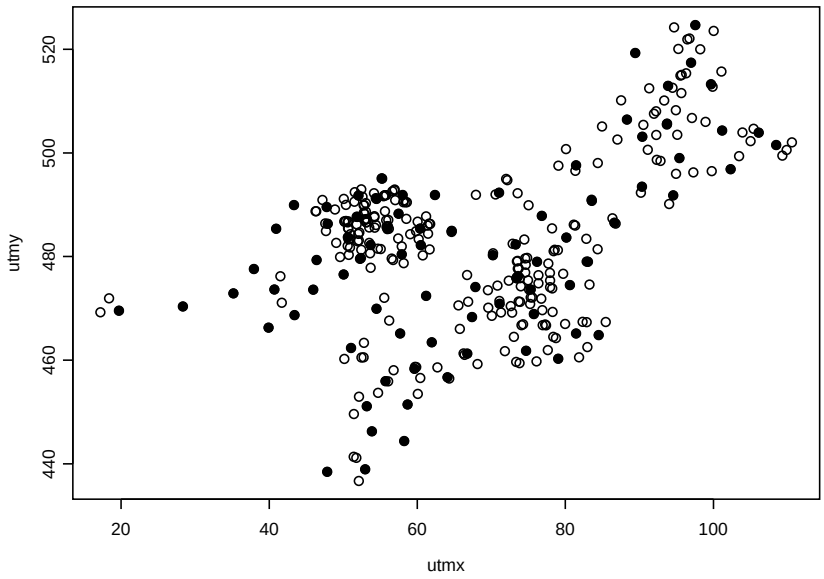

Figure 3: Location of knots for the bivariate radial spline function on the UTM coordinates.

the bootstrap procedure both with and without including the matrix $\boldsymbol{Q}$ discussed in that section, and found almost no difference between them.

As noted above, other mean model specifications were also evaluated, including the addition of linear terms for the North/South and East/West spatial coordinates and a quadratic term for elevation. None of those terms were found to be statistically significant. The coefficient for the intercept in Table 1 is also not statistically significant, but it was not removed from the model as it was significant in some of the fits with different random effects specifications (see below).

Figure 4 shows a map with the small area predictions $\widehat{y}_t$ for all HUCs. Compared to the map in Figure 2, the small area estimation map is smoother and also contains values in all HUCs, offsetting some of the limitations of the original data. One noticeable difference between the HUC mean map and the small area map is that the smallest value in the latter is negative. ANC values can indeed be negative, and the dataset contains 39 negative observations (out of 551), with a smallest observation of -72.2. Hence, while the small area predicted value of -37.6 indeed falls outside of the range of the HUC means, it is well within the range of the observed data.

Even though both the spline and the small area random effects appear to be highly statistically significant as measured by the bootstrap-based likelihood ratio test, it is still of interest to further investigate what the practical impact is of including both random effects relative to simpler models that only include one of them. Table

Fixed effects		
Parameter	$\widehat{\beta}$	P-value
Intercept	228.6	0.87
Elevation	-0.814	$< .001$
Random effects		
Parameter	$\widehat{\sigma}$	P-value
Spline	71.2	$< .001$
HUC	365.7	$< .001$
Errors	179.5	$< .001$

Table 1: Parameter estimates for penalized spline small area estimation model for Northeastern Lakes data.

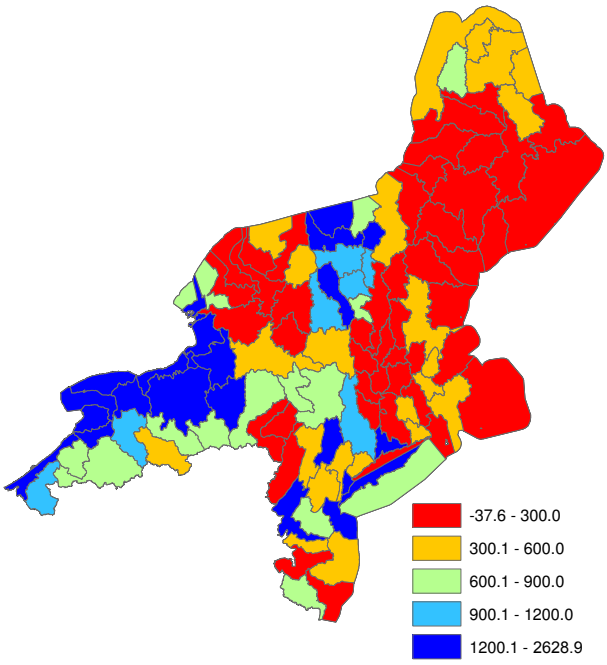

Figure 4: Map of model predicted mean ANC for all HUCs.

		HUC	
		yes	no
Spline	yes	0.98 / 7755	0.88 / 7894
	no	0.99 / 7968	0.02 / 8497

Table 2: Comparison of correlation / AIC values between HUC model predictions and averages of the sample observations in the HUCs for inclusion and exclusion of random effect terms in model.

2 shows the correlations between the $\widehat{y}_t$ and averages of the sample observations in the HUCs for four cases, depending on whether each of the two random effects is included in the model or not, as well as the corresponding AIC values. The highest correlation is achieved by the model with a HUC random effect but no spline random effect, while the smallest AIC is attained by the model with both random effect. The model with a spline random effect but no HUC random effect achieves an AIC that is lower than that of the model with both random effects reversed, even though its correlation is slightly lower. All three models with at least one random effect outperform the model with only fixed effects.

Judging by these criteria, the models with either the HUC or the spline random effect, but not both, achieve small area predictions that are roughly as good as the model with both random effects. Such model fitting criteria provide an incomplete view of the usefulness of the model, however. In Figures 5 and 6, we plot the HUC predictions obtained by the full model against those for the models with single random effects for a further comparison.

Figure 5 shows that the HUC-only model and the model with both random effects result in similar predictions for HUCs containing sample observations, but dramatically different predictions for the HUCs without observations. Relative to the HUC-only model, the addition of the spatial spline term appears to improve model predictions for these "empty" HUCs, by borrowing strength from neighboring observations located in different HUCs. In contrast, a HUC-only model predicts a HUC effect of 0 in those empty HUCs, so that only the fixed linear part of the model is used in prediction. This likely improvement in model fit is not captured by either AIC or correlation, so that it is not reflected in summary statistics such as those in Table 1.

In Figure 6, differences between the spline-only model and that with both random effects are not as clear, but some large deviations from the 45-degree line are still

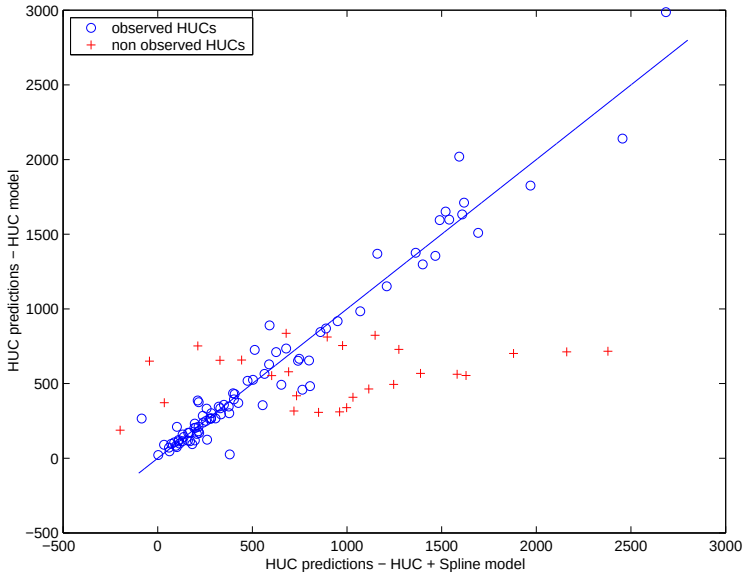

Figure 5: Comparison of HUC predictions for model with both random effects and model with HUC random effect only (solid line is 45-degree line).

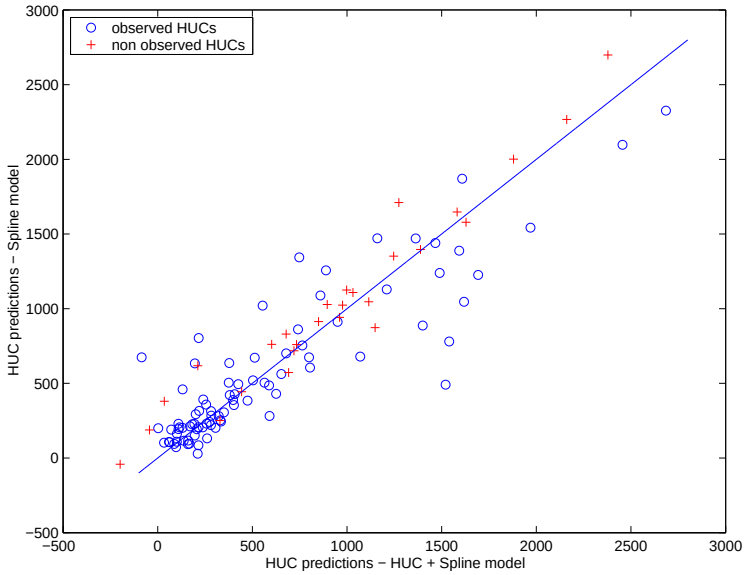

Figure 6: Comparison of HUC predictions for model with both random effects and model with spline random effect only (solid line is 45-degree line).

present. Differences between both fits can be explained by the fact that both models attempt to fit different "targets": whereas the spline-only model predicts a smooth spatial trend for the region of interest, the model with both effects predicts small area HUC means of the form (6), which include both a smooth and a HUC-specific effect. Since the goal of small area estimation is to capture features that might be unique to lakes in particular HUCs, a small area estimation model that makes it possible to do so when sufficient HUC-specific data are available is clearly preferred. In Figure 6, this is illustrated by the fact that the predictions for "empty" HUCs tend to be closer to the 45-degree line than the predictions for the remaining HUCs.

Acknowledgments

The work of Opsomer, Breidt, and Ranalli was developed under STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA). This paper has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report.

References

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association. 83*, 28–36.

Claeskens, G. (2004). Restricted likelihood ratio lack-of-fit tests using mixed spline models. *Journal of the Royal Statistical Society, Series B 66*, 909–926.

Coull, B. A., D. Ruppert, and M. P. Wand (2001). Simple incorporation of interactions into additive models. *Biometrics 57*, 539–545.

Coull, B. A., J. Schwartz, and M. P. Wand (2001a). Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics 2*, 337–349.

Crainiceanu, C. and D. Ruppert (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B 66*, 165–185.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Stat. Science 11*(2), 89–121.

Ferguson, T. S. (1996). *A course in large sample theory*. New York: Chapman & Hall/CRC.

Ghosh, M. and J. Rao (1994). Small area estimation: an appraisal. *Statistical Science 9*, 55–93.

Jiang, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statistica Sinica 8*, 861–885.

Larsen, D. P., T. M. Kincaid, S. E. Jacobs, and N. S. Urquhart (2001). Designs for evaluating local and regional scale trends. *Bioscience 51*, 1049–1058.

Liang, H. (2003). Penalized spline for nonparametric regression analysis of longitudinal data. unpublished manuscript.

McCulloch, C. and S. Searle (2001). *Generalized, Linear and Mixed Models*. New York: Wiley.

Messer, J. J., R. A. Linthurst, and W. S. Overton (1991). An EPA program for monitoring ecological status and trends. *Environmental Monitoring and Assessment 17*, 67–78.

Nychka, D., P. Haaland, M. O'Connel, and S. Ellner (1998). FUNFITS, data analysis and statistical tools for estimating functions. In D. Nychka, W. Piegorsch, and L. H. Cox (Eds.), *Case studies in environmental statistics*, pp. 159–179. New York: Springer.

Parise, H., D. Ruppert, L. Ryan, and M. P. Wand (2001). Incorporation of historical controls using semiparametric mixed models. *Applied Statistics 50*, 31–42.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

Ruppert, R., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics 18*, 223–249.

Wetzel, R. G. (1975). *Limnology*. Philadelphia: W.B. Saunders Company.

Zheng, H. and R. J. A. Little (2003). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. unpublished manuscript, submitted to *Survey Methodology*.