

Nonparametric Small Area Estimation Using Penalized Spline Regression

J. D. Opsomer

Colorado State University*

G. Claeskens

Katholieke Universiteit Leuven

M. G. Ranalli

Universita' degli Studi di Perugia

G. Kauermann

Universität Bielefeld

F. J. Breidt

Colorado State University

11th September 2007

Abstract

This article proposes a 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 nonparametric small area estimation problem as a mixed effect model regression. The resulting model is readily fitted using existing model fitting approaches such as restricted maximum likelihood. We present theoretical results on the prediction mean squared error of the proposed estimator and on likelihood ratio tests for random effects, and we propose a simple nonparametric 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.

*Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA; jopsomer@stat.colostate.edu.

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-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) review the most commonly used types of estimators used by survey statisticians, including synthetic and composite estimators, mixed model prediction, and empirical and hierarchical Bayesian approaches. The “canonical” small area estimation model is a linear mean model for the data and a random effect for the small areas, with both masked by an additional amount of noise due to not having sampled the complete small area. Both the random effect and the noise are assumed to be independent realizations from underlying distributions. The response variable can either be observed at the small area level, or at a smaller unit or respondent level. Fay and Herriot (1979) studied the *area-level* model and proposed an empirical Bayes estimator for that case. Battese et al. (1988) considered the *unit-level* model and constructed an empirical best linear unbiased predictor (EBLUP) for the small area means. Numerous extensions to this setup have been considered in the literature, including for data that follow various generalized linear models and have more complicated random effects structures. Rao (2003) provides a good overview of the available estimation methods, and Jiang and Lahiri (2006) review the theoretical development of mixed model estimation in the small area context. The extension we are considering here is to incorporate nonparametric regression models in small area estimation, which we will do for the unit-level case.

In principle, a nonparametric model might have significant advantages compared to 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. As will be made more specific below, the two concepts underlying P-splines are the replacement of the fully nonparametric mean trend by a highly parametrized function form, and the imposition of penalty to ensure that the parameter estimators achieve good statistical properties. Hence, even though penalized spline regression is most often referred to as a nonparametric method, it really represents a flexible class of parametric methods based on linear models. In the current article, we exploit the close connection between P-splines and linear mixed models (see Wand, 2003) to show how to incorporate a nonparametric mean function specification into existing small area estimation approaches.

The ability to combine nonparametric regression and mixed model regression with P-splines has been used in other contexts. Parise et al. (2001), Coull et al. (2001) and Coull et al. (2001a) 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 (2004) 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 proposed method is also related to linear mixed model approaches in which complex data structures are captured through more sophisticated random effects structures. Related approaches include, for instance, Clayton and Kaldor (1987), who proposed a model in which the small area random effects are correlated, and Ghosh et al. (1998), who used a prior distribution for the small area effect that includes spatial correlation between small areas. Further related models are described in Rao (2003, Ch. 8). In these models, a simple mean model is supplemented by a random effect specification that makes it possible to capture relationships between neighboring small areas. While the P-spline model can also be used to incorporate spatial proximity effects (as will be done in the application considered later in this article), the method can be applied more generally to modeling situations in which the relationship between dependent and independent variables cannot be properly captured by a simple parametric structure.

The goal of the article is to demonstrate how nonparametric regression and related inference methods can be incorporated into the various components of small area estimation and inference, using as a case study a survey of lake water quality vari-

ables. In Section 2, we briefly review penalized spline regression and show how to incorporate it in small area estimation. Section 3 presents theoretical properties of the proposed method, including the prediction mean squared error of the small area estimates and an estimator for that quantity. We also discuss likelihood ratio testing for the significance of the spline term and the small area random effect, and we propose a simple bootstrap method that is easy to implement and is applicable to both mean squared error estimation and testing. Throughout this section, our main emphasis is on extending and/or applying existing approaches, rather than developing new theoretical results.

Section 4 contains the case study, based on data from 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, and use the bootstrap approach to do model inference. We also conduct a limited simulation study to evaluate the validity of the bootstrap approach in this context.

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_ε^2 . 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

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

Here p is the degree of the spline, $(x)_+^p$ denotes the function $x^p \mathbf{I}_{\{x > 0\}}$, $\kappa_1 < \dots < \kappa_K$ is a set of fixed *knots* and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)'$ are the coefficient vectors for the “parametric” and the “spline” portions of the model, respectively. Provided the knot locations are sufficiently spread out over the range of x and 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 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 $m(\cdot; \boldsymbol{\beta}, \boldsymbol{\gamma}) - m(\cdot; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$. Ruppert (2002) provides simulation-based evidence that this lack-of-fit error is indeed negligible in the univariate nonparametric regression case.

The spline function (1) uses the *truncated polynomial spline basis* $\{1, x, \dots, x^p, (x - \kappa_1)_+^p, \dots, (x - \kappa_K)_+^p\}$ to approximate the function m_o . 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.

Following the recommendations in Ruppert (2002), the knots are often at equally spaced quantiles of the distribution of the covariate and K is taken to be large relative to the size of the dataset. A typical knot choice for univariate x would be 1 knot every 4 or 5 observations, with a maximum number of 35-50. For multivariate regression problems, other approaches are recommended to “spread out” the knots over the covariate space, and we will return to this in Section 4. In both situations, the model (1) is potentially over-parameterized and difficult to fit. This issue is avoided by putting a *penalty* on the magnitude of the spline parameters $\boldsymbol{\gamma}$. For a given dataset $\{(x_i, y_i) : i = 1, \dots, n\}$, this is done by defining the regression estimators as the minimizers over $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of

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

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 vector 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. the nested error regression model of 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 for which estimates are to be constructed. Define d_{it} as the indicator taking value of 1 if observation i is in small area t and 0 otherwise, and let $\mathbf{d}_i = (d_{i1}, \dots, d_{iT})'$. We also define $\mathbf{Y} = (y_1, \dots, y_n)'$,

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

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon} \quad (2)$$

where

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

and each of the random components is assumed independent of the others. The model (2) includes the spline function, which can be thought of as a nonparametric mean function specification, and the small area random effects $\mathbf{D}\mathbf{u}$. For the purpose of fitting this model and using the appropriate amount of smoothing for the spline, it is convenient to continue to treat $\mathbf{Z}\boldsymbol{\gamma}$ as a random effect term, so that $\text{Var}(\mathbf{Y}) \equiv \mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_\gamma\mathbf{Z}' + \mathbf{D}\boldsymbol{\Sigma}_u\mathbf{D}' + \boldsymbol{\Sigma}_\varepsilon$.

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

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

and the predictors

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

are optimal among all linear estimators/predictors.

For a given small area t , we are interested in predicting

$$\bar{y}_t = \bar{\mathbf{x}}_t \boldsymbol{\beta} + \bar{\mathbf{z}}_t \boldsymbol{\gamma} + u_t, \quad (6)$$

where $\bar{\mathbf{x}}_t, \bar{\mathbf{z}}_t$ 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, which incorporates area-level unmodeled random variation. Both $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{z}}_t$ are assumed known. Note that \bar{y}_t is not generally equal to the true mean of the y_i in the small area, 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, $u_t = \bar{\mathbf{d}}_t \mathbf{u} = \mathbf{e}_t \mathbf{u}$, where \mathbf{e}_t is a vector with 1 in the t th position and 0s everywhere else. As a predictor of \bar{y}_t , we therefore use

$$\hat{y}_t = \bar{\mathbf{x}}_t \hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_t \hat{\boldsymbol{\gamma}} + \mathbf{e}_t \hat{\mathbf{u}}, \quad (7)$$

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

If the variances are unknown, a commonly used approach in mixed model regression is to use so-called EBLUP versions of (4), (5) and (7), which are constructed by replacing $\sigma_\gamma^2, \sigma_u^2, \sigma_\varepsilon^2$ 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, lme() in S-Plus and R, or by using programs specifically written for penalized spline regression such as the SemiPar package in R.

3 Theoretical Properties

3.1 Prediction Mean Squared Error

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

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

Then,

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

with $\mathbf{c}_t = \bar{\mathbf{x}}_t - \bar{\mathbf{w}}_t \boldsymbol{\Sigma}_w \mathbf{W}' \mathbf{V}^{-1} \mathbf{X}$. 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

$$E(\hat{y}_t - \bar{y}_t)^2 = \mathbf{c}_t'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{c}_t + \bar{\mathbf{w}}_t'\boldsymbol{\Sigma}_w(\mathbf{I} - \mathbf{W}'\mathbf{V}^{-1}\mathbf{W}\boldsymbol{\Sigma}_w)\bar{\mathbf{w}}_t'. \quad (9)$$

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

If the variances of the random effects are estimated from the data, the resulting EBLUP version of (8) is

$$\tilde{y}_t - \bar{y}_t = \hat{\mathbf{c}}_t'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \bar{\mathbf{w}}_t'(\hat{\boldsymbol{\Sigma}}_w\mathbf{W}'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\omega}) \quad (10)$$

with $\hat{\mathbf{c}}_t = \bar{\mathbf{x}}_t - \bar{\mathbf{w}}_t'\hat{\boldsymbol{\Sigma}}_w\mathbf{W}'\hat{\mathbf{V}}^{-1}\mathbf{X}$, using REML estimators for the unknown variance components in \mathbf{V} and $\boldsymbol{\Sigma}_w$. Expression (9) is no longer equal to the MSE of the prediction errors for the EBLUP, and a substantial literature exists on approximations and estimators for the MSE of small area estimators for both area-level and unit-level models. In the case of small area estimation with a linear mean model and independent variance components, Prasad and Rao (1990) extended the results of Kackar and Harville (1984) to derive a second-order approximation for the prediction MSE (PMSE) as well as an estimator for the PMSE that is correct up to second order. Datta and Lahiri (2000) later extended their results for the case of REML estimation of the variance components, and Das et al. (2004) further expanded it to encompass more general linear mixed models. Two important characteristics of these methods are (i) that the approximations to the PMSE include the effect of the estimation of the random effect parameters, and (ii) that the PMSE estimators need to include a bias correction term in order to be consistent for the PMSE.

For the case with a spline-based random component, we have the result as formulated in the following theorem, which states a second order approximation to the PMSE of the EBLUP, together with its estimator, also correct to the second order. Hence, the spline-based small area estimation approach achieves the same two characteristics as the above methods. This result and the method of proof are extensions of Das et al. (2004) to the case of a spline-based random effect. However, it should be noted that because of the structure of the variance-covariance matrix induced by the spline random component, the results of Das et al. (2004) do not apply directly. First we make the following definitions. Let $\boldsymbol{\sigma}^2 = (\sigma_\gamma^2, \sigma_u^2, \sigma_\varepsilon^2)$. Let \mathcal{S} be a matrix with rows $\mathcal{S}_j = \bar{\mathbf{w}}_t' \left(\frac{\partial \boldsymbol{\Sigma}_w}{\partial (\sigma^2)_j} \mathbf{W}' \mathbf{V}^{-1} + \boldsymbol{\Sigma}_w \mathbf{W}' \frac{\partial \mathbf{V}^{-1}}{\partial (\sigma^2)_j} \right)$, $j=1,2,3$, where $\frac{\partial \boldsymbol{\Sigma}_w}{\partial (\sigma^2)_1} \equiv \frac{\partial \boldsymbol{\Sigma}_w}{\partial \sigma_\gamma^2} =$

$\text{diag}(I_K, 0_T)$, $\frac{\partial \boldsymbol{\Sigma}_w}{\partial (\sigma^2)_2} \equiv \frac{\partial \boldsymbol{\Sigma}_w}{\partial \sigma_u^2} = \text{diag}(0_K, I_T)$, $\frac{\partial \boldsymbol{\Sigma}_w}{\partial (\sigma^2)_3} \equiv \frac{\partial \boldsymbol{\Sigma}_w}{\partial \sigma_\varepsilon^2} = 0_{K+T}$ and $\frac{\partial \mathbf{V}^{-1}}{\partial (\sigma^2)_j} = -\mathbf{V}^{-1} \mathbf{B}_j \mathbf{V}^{-1}$ with $\mathbf{B}_1 = \mathbf{Z}\mathbf{Z}'$, $\mathbf{B}_2 = \mathbf{D}\mathbf{D}'$ and $\mathbf{B}_3 = \mathbf{I}_n$. Further, the 3×3 matrix \mathcal{I} , the Fisher information matrix with respect to $\boldsymbol{\sigma}^2$, contains elements $\mathcal{I}_{ij} = \frac{1}{2} \text{tr}(\mathbf{P}\mathbf{B}_i \mathbf{P}\mathbf{B}_j)$, where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}'\mathbf{V}^{-1}$.

Theorem 3.1 *Assume that there exists a value $\delta > 1$ such that $E(|y_i|^{2\delta})$ is bounded, that the true variance components $\boldsymbol{\sigma}^2 = (\sigma_\gamma^2, \sigma_u^2, \sigma_\varepsilon^2)$ are positive, that the largest eigenvalue of \mathbf{V} is $O(L_n)$, where $L_n = o(\sqrt{n})$, and that the number of small areas $T = O(n)$ and the number of knots K is fixed. Then, the prediction mean squared error of the EBLUP predictor in (10) is given by*

$$PMSE(\tilde{y}_t) \equiv E(\tilde{y}_t - \bar{y}_t)^2 = E(\hat{y}_t - \bar{y}_t)^2 + \text{tr}(\mathbf{S}\mathbf{V}\mathbf{S}'\mathcal{I}^{-1}) + o(L_n^2/n). \quad (11)$$

If we assume that $L_n = O\{(\log \sqrt{n})^a\}$ for any positive a , then the estimator of the quantity in (11),

$$\begin{aligned} \widehat{PMSE}(\hat{y}_t) &= \hat{\mathbf{c}}_t' (\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \hat{\mathbf{c}}_t + \bar{\mathbf{w}}_t' \hat{\boldsymbol{\Sigma}}_w \left(\mathbf{I} - \mathbf{W}' \hat{\mathbf{V}}^{-1} \mathbf{W} \hat{\boldsymbol{\Sigma}}_w \right) \bar{\mathbf{w}}_t \\ &\quad + 2(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \hat{\mathbf{S}}' \hat{\mathcal{I}}^{-1} \hat{\mathbf{S}} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \end{aligned}$$

inserting REML estimators for unknown variance components in \mathcal{S} and \mathcal{I} , is second order correct. That is, $E(\widehat{PMSE}) = PMSE + o(L_n^2/n)$.

The proof is given in the Appendix.

3.2 Testing for small area effects and non-linearities

In the model, there are two main sources of variability in addition to the pure error term: (i) the small area effects, and (ii) the deviation from the parametric p th degree polynomial model, as accounted for by the spline terms. 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. We therefore propose a likelihood ratio test (or restricted likelihood ratio test) for testing the presence of small area effects. 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 log 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 p th

degree polynomial, $H_{0,\gamma} : \sigma_\gamma^2 = 0$ versus $H_{a,\gamma} : \sigma_\gamma^2 > 0$ is denoted by \mathcal{L}_γ . 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 $\sigma_\gamma^2 > 0$. The corresponding (restricted) log likelihood value is denoted $\mathcal{L}_{\gamma,u}$.

One-sided testing has a long history, going back to Chernoff (1954). An important reference for an asymptotic study of likelihood ratio tests under boundary constraints is Self and Liang (1987), applicable to independent and identically distributed data. Stram and Lee (1994) consider tests on variance components in a longitudinal mixed linear model, and Vu and Zhou (1997) provide general theoretical results. The overview paper by Sen and Silvapulle (2002) contains more information on the particular type of asymptotic distribution obtained in such testing problems, as well as results for the related Wald and score tests. Because of the special characteristics of the spline random effect, the results of these authors do not apply directly to the estimator we are considering here. Instead, we will extend existing results for testing spline random effects to the case where there is an additional small area effect.

Depending on the assumptions one is willing to make, it is possible to theoretically derive either an asymptotic distribution or an exact one for the above likelihood ratio statistics. If we assume that both T and K tend to infinity such that $T = o(n)$ and $K = o(n)$, together with some regularity conditions on the design matrices and the moments of the random variables, and we define $\boldsymbol{\lambda} = (\lambda_\gamma, \lambda_u) = (\sigma_\gamma^2/\sigma_\epsilon^2, \sigma_u^2/\sigma_\epsilon^2)$, with $\boldsymbol{\lambda}_0$ its value under the null hypothesis for any of the three hypotheses, we then have the following result: 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}_γ) 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 $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, which would correspond to the result one would expect under the theory developed for the independent case by Self and Liang (1987). The proof of this result follows the same line of arguments as that of Theorem 2 of Claeskens (2004), with the difference that only one variance component is set to zero. A similar result is obtained for the joint testing problem, with the asymptotic distribution a mixture of χ_0^2, χ_1^2 and χ_2^2 .

These asymptotic approximations are very easy to compute, but as discussed in Crainiceanu and Ruppert (2004) in the context of testing for the presence of the spline components, they do not perform satisfactorily in practice. Moreover, we prefer not to make the assumption on the number of spline components K growing to infinity, since this would be in conflict with the assumptions of Theorem 3.1.

An alternative to using the asymptotic distribution is to proceed along the same lines as in Crainiceanu et al. (2005) for testing polynomial regression models under the assumption of Gaussian random variables, and to obtain an *exact* restricted likelihood ratio test via spectral decompositions. Our situation differs from the setting considered in that article because of the presence of the small area effect.

The algorithm to simulate the finite sample distribution is a generalization of that given for the case of a single variance component in Crainiceanu and Ruppert (2004). While the principle is similar, the case with two variance components poses additional challenges. The most important one is that the required eigenvalues for this algorithm do not have an explicit form in terms of $(\lambda_\gamma, \lambda_u)$. Therefore, for each evaluation of the likelihood at given values of $(\lambda_\gamma, \lambda_u)$, one would need to recompute these eigenvalues. This difficulty, combined with the fact that the grid search is in more than one dimension, make this procedure cumbersome to use in practice. In the next section, we will discuss a more practical bootstrap-based method that is applicable for both estimation of the prediction mean squared error and testing for the random components.

3.3 Bootstrapping small area and local effects

Bootstrap replicate observations are generated as

$$\mathbf{Y}^* = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{D}\mathbf{u}^* + \boldsymbol{\varepsilon}^*, \quad (12)$$

where $\boldsymbol{\gamma}^*$, \mathbf{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, normal distributions would often be used for this purpose. Butar and Lahiri (2003) consider parametric bootstrap estimation of the prediction mean squared error assuming Gaussian distributions. Lahiri (2003) reviews other proposals for prediction mean squared error estimation using parametric bootstrap. Recent work on parametric bootstrap methods includes Hall and Maiti (2006), who propose a double bootstrap procedure to construct a second order correct estimator of the PMSE.

One drawback of the parametric bootstrap approach is that it could lead to biased inference if these distributions are misspecified. Pfeiffermann and Glickman (2004) propose a nonparametric bootstrap for the Fay-Herriot model, which starts from

an asymptotic approximation to the small area prediction mean squared error as in (11) and uses standardized resampled residuals to estimate individual terms of that approximation. The approach we propose here also uses resampling, but there are two main differences between our approach and that of Pfeiffermann and Glickman (2004). First, we do not target the asymptotic approximation of the PMSE and second, instead of using the residuals themselves, we split them into components corresponding to the random effects and model errors. An advantage of our approach is that it makes it possible not only to estimate the PMSE, but also to obtain bootstrap distributions for the likelihood ratio test statistics, as will be described below.

Assuming that the variances σ^2 are known, we start from the BLUP predictors $\hat{\gamma}$ and $\hat{\mathbf{u}}$ obtained in (5), and denote $\mathbf{H} = \mathbf{I} - \mathbf{V}\mathbf{P}$. In this case, the variance of $\hat{\gamma}$ is

$$\text{Var}(\hat{\gamma}) = \sigma_\gamma^4 \mathbf{Z}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{H})\mathbf{Z}$$

so that we need to correct $\hat{\gamma}$ before being able to draw bootstrap replicates γ^* with the proper second moment. We therefore replace the $\hat{\gamma}$ by

$$\tilde{\gamma} = (\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{H})\mathbf{Z})^{-1/2}\hat{\gamma}/\sigma_\gamma \quad (13)$$

before resampling. The same reasoning leads to

$$\tilde{\mathbf{u}} = (\mathbf{D}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{H})\mathbf{D})^{-1/2}\hat{\mathbf{u}}/\sigma_u. \quad (14)$$

Finally, to generate estimated errors, we start from

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\boldsymbol{\gamma}} - \mathbf{D}\hat{\mathbf{u}} = \boldsymbol{\Sigma}_\varepsilon\mathbf{V}^{-1}(\mathbf{I} - \mathbf{H})\mathbf{Y},$$

so that $\text{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma_\varepsilon^4\mathbf{V}^{-1}(\mathbf{I} - \mathbf{H})$ and hence,

$$\tilde{\boldsymbol{\varepsilon}} = (\mathbf{V}^{-1}(\mathbf{I} - \mathbf{H}))^{-1/2}\hat{\boldsymbol{\varepsilon}}/\sigma_\varepsilon. \quad (15)$$

Bootstrap resampling is done with replacement from $\tilde{\gamma}$, $\tilde{\mathbf{u}}$ and $\tilde{\boldsymbol{\varepsilon}}$ after centering to obtain zero-mean random components. The resulting bootstrap distributions of γ^* , \mathbf{u}^* and $\boldsymbol{\varepsilon}^*$ will therefore have variances $\boldsymbol{\Sigma}_\gamma^* = \tilde{\boldsymbol{\gamma}}'\tilde{\boldsymbol{\gamma}}/K\mathbf{I}_K$, $\boldsymbol{\Sigma}_u^* = \tilde{\mathbf{u}}'\tilde{\mathbf{u}}/T\mathbf{I}_T$ and $\boldsymbol{\Sigma}_\varepsilon^* = \tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}/n\mathbf{I}_n$, respectively. In the case of known variance components considered so far, these bootstrap variances can be expected to converge to $\boldsymbol{\Sigma}_\gamma$, $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\varepsilon$ under mild conditions, but we will not investigate this further here.

In practice, the variance components are unknown and are estimated from the data. This above procedure is therefore “naive” in the sense that it does not account for

the estimation of the variance parameters, and hence should be viewed as a BLUP bootstrap which will generally not be second order correct for the estimation of the PMSE. As will be further illustrated in Section 4, the error of the parameter estimation of σ^2 is often very small relative to the prediction error, especially in applications with unit-level (as opposed to small-area-level) models, so that a bootstrap approach that ignores estimation uncertainty is likely to perform almost as well in many cases as a more complicated procedure that attempts to incorporate that portion of the error. Therefore, in the bootstrap procedure we implemented, we used the expressions (13), (14) and (15) after replacing the variance components σ^2 by their REML estimators $\hat{\sigma}^2$. Similarly, we also found that the matrix \mathbf{H} , which accounts for the estimation of β , had virtually no effect on the adjustments.

In Section 4, we discuss a limited simulation study that assesses the appropriateness of the nonparametric bootstrap in the context of the application. Research in a full EBLUP version of this simple bootstrap would certainly be warranted. In principle, it would be possible to improve the second moment corrections in (13)–(15) by extending the results of Theorem 3.1, since each of the elements of $\hat{\gamma}$ and $\hat{\mathbf{u}}$ is of the same form as \tilde{y}_t (for different $\tilde{\mathbf{w}}_t$). Alternatively, a “plug-in” nonparametric bootstrap like that of Pfeiffermann and Glickman (2004) could be derived. We do not pursue this further here.

Once the bootstrap random components and errors are generated, bootstrap observations \mathbf{Y}^* are constructed using (12). Drawing B bootstrap samples obtained in this manner, the PMSE for the small areas are estimated by

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

where superscript b indexes the bootstrap samples. The above bootstrap approach can also be applied in the testing context. We illustrate the use of the bootstrap procedure in approximating the distribution of the likelihood ratio statistic for the case $H_{0,u} : \sigma_u^2 = 0$. 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

$$\mathbf{Y}^* = \mathbf{X}\hat{\beta} + \mathbf{Z}\gamma^* + \varepsilon^*,$$

where γ^* and ε^* are generated as discussed above. For each bootstrap replicate sample $b = 1, \dots, B$, we fit the restricted and the full model and obtain \mathcal{L}_u^{*b} . The significance of \mathcal{L}_u is then evaluated against the empirical distribution of \mathcal{L}_u^{*b} .

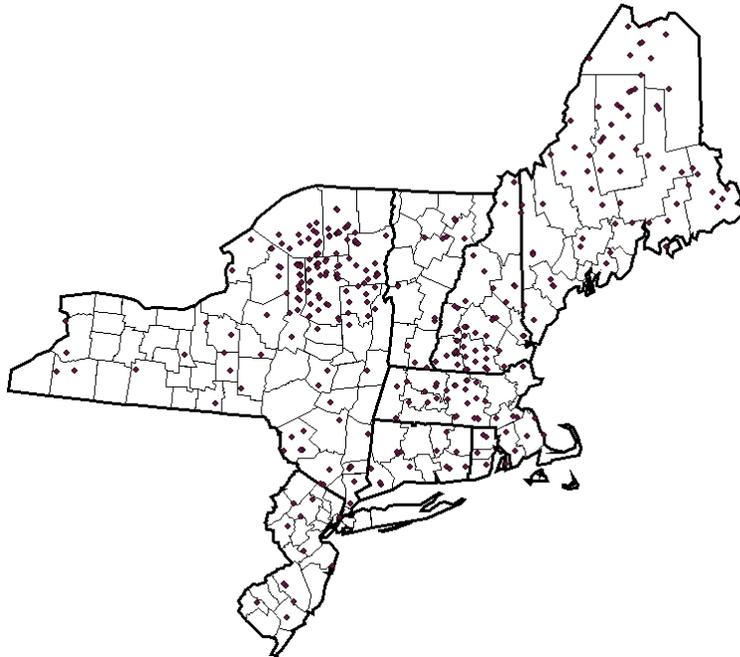


Figure 1: Locations of sampled lakes in Northeastern U.S.

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) within the region of interest. HUCs represent a nested subdivision of all U.S. land based on hydrological features, and are frequently used in delineating areas of analysis in surveys of natural resources. 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.

The goal of the analysis is to identify HUCs of concern within the region, based

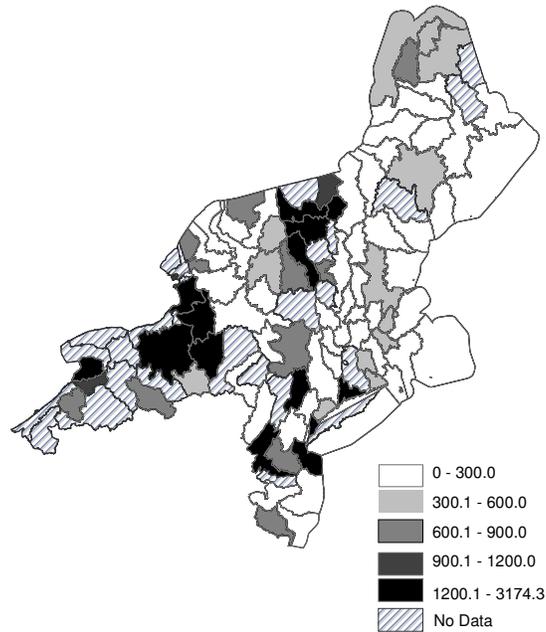


Figure 2: Hydrologic Unit Code (HUC) small areas within Northeastern U.S. region, with average ANC computed in all small areas containing sample observations.

on the results from the survey. HUCs are a meaningful subdivision of the region, because HUC boundaries typically follow watershed drainage areas, and lakes in a HUC are expected to be influenced by the same hydrological and associated features. Hence, lakes in close geographical proximity but located in different HUCs are expected to be less similar than two lakes in the same HUC. At the same time, factors affecting ANC such as acid deposition and soil characteristics cut across HUCs, so that overall spatial trends are also likely to be useful in predicting ANC. Therefore, a HUC prediction model that captures spatial trends and allows for HUC-specific effects has the potential to capture most of the interesting patterns in the data, and will be explored in this section.

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

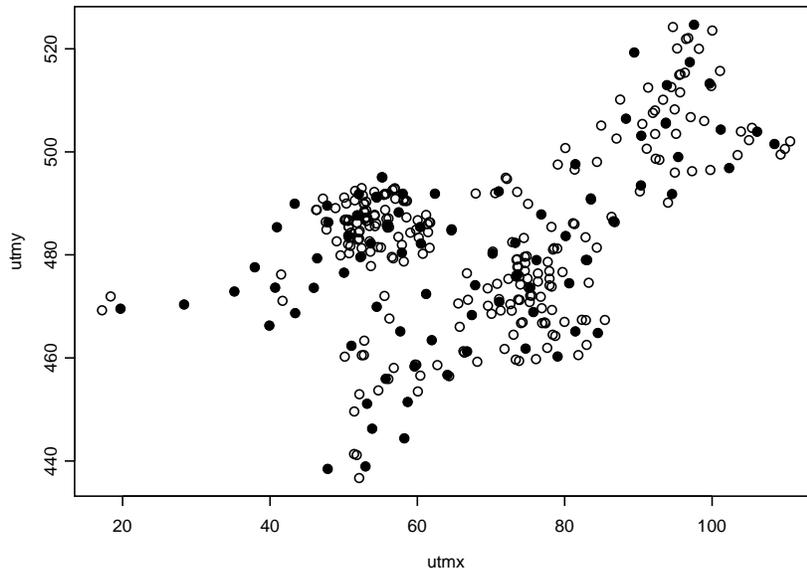


Figure 3: Lake locations (open circles) and knot locations of the bivariate radial spline function on the UTM coordinates (solid circles).

that a bivariate spline on the UTM coordinates and a linear term for elevation provided the best model fit. We 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 follow Ruppert et al. (2003, p.253) in using a transformed *radial basis*, defined as

$$\mathbf{Z} = [C(\mathbf{x}_i - \boldsymbol{\kappa}_k)]_{\substack{1 \leq i \leq n \\ 1 \leq k \leq K}} [C(\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'})]_{1 \leq k, k' \leq K}^{-1/2}, \quad (16)$$

where $C(\mathbf{r}) = \|\mathbf{r}\|^2 \log \|\mathbf{r}\|$, $\mathbf{x}_i = (x_{1i}, x_{2i})$ denotes the geographical coordinates for observation i and $\boldsymbol{\kappa}_k, k = 1, \dots, K$ are spline knots. The second matrix on the right-hand side of (16) applies a linear transformation to the radial basis functions in the first matrix, and is recommended by Ruppert et al. (2003) as a way to make the radial spline behave approximately like a thin-plate spline.

Knot selection for spatial applications is discussed in Ruppert et al. (2003, p.255). Since it is no longer possible to place knots at the quantiles of the covariate as in the univariate case described in Section 2, the use of space filling designs is recommended to ensure coverage of the covariate space as well as parsimony in the number of knots. We used the space-filling algorithm implemented in the `cover.design()` function

Fixed effects		
Parameter	$\hat{\beta}$	P-value
Intercept	228.6	0.87
Elevation	-0.814	< .001
Random effects		
Parameter	$\hat{\sigma}$	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.

in the `FUNFITS` package for `S-plus` (Nychka et al. 1998) for this purpose. Figure 3 shows the locations of the $K = 80$ knots selected by this method.

The ANC small area model can now be written as in (2) with variance components (3). That model includes \mathbf{Y} for the ANC observations, \mathbf{X} for the matrix containing an intercept and the linear elevation term, \mathbf{Z} as in (16) for the spatial locations, and \mathbf{D} for the matrix of indicators for the HUCs. This model is fitted using REML as implemented in `lme()` in `S-plus`, and calculations take approximately 6.6s for a single fit on a Pentium 1.6GHz Intel laptop. The parameter estimates and corresponding P-values are shown in Table 1. The P-values are computed using the bootstrap procedure described in Section 3.3 and a bootstrap sample size $B = 5000$. Since the P-values are all far from the customary cut-off value of 0.05, we felt this was a sufficiently large number of bootstrap replicates to be able to determine the significance (or lack thereof) of the parameter estimates.

The left plot in Figure 4 shows a map with the small area predictions \hat{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 EBLUP map is that the smallest value in the latter is negative. ANC values can 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. This map is a combination of a linear elevation effect, a smooth spatial trend captured by the spline random effect, and a HUC-specific random deviation. The right plot

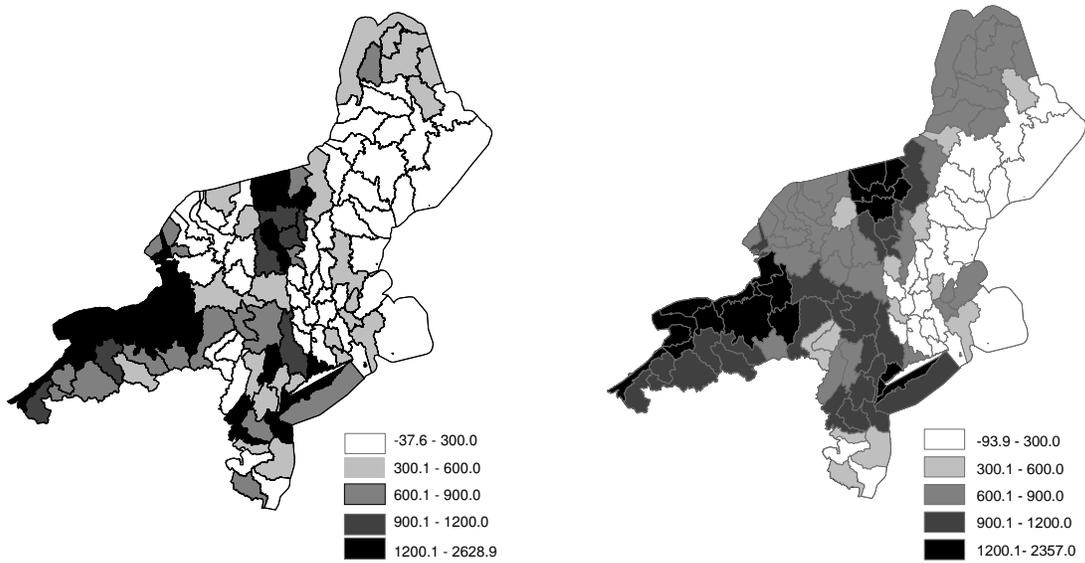


Figure 4: Left: map of predicted mean ANC for all HUCs. Right: spline component predictions.

in Figure 4 displays the prediction of the spatial spline surface averaged for each HUC. A smoothly varying spatial trend is clearly visible, with high values in the West and low values in both the South and the East.

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). We also investigated the effect of the number of knots, and repeated the analysis with K varying from 40 to 100. We applied AIC to provide guidance on the knot selection, and found that the AIC decreases over that range, with the rate of decrease steeper between 40 and 80 and much less so between 80 and 100. While the parameter estimates varied somewhat over that range of values, the overall fits remained similar and in particular the significance of the parameters did not change. Hence, we are comfortable with $K = 80$ as a suitable compromise between model parsimony and goodness of fit.

In order to estimate the uncertainty associated with the predictions in Figure 4, we computed estimated PMSE for the small areas using both the asymptotic approximation from Theorem 3.1 and the nonparametric bootstrap. As noted in Section 3.3, the bootstrap method ignores the uncertainty associated with parameter es-

timation. To assess the appropriateness of omitting this source of variability, we first computed the estimated PMSE according to Theorem 3.1 for each of the HUCs and compared the relative size of the three PMSE components. Averaged over the HUCs, the prediction component represents 96.9% of the estimated PMSE, the estimation of the mean model parameters 1.1% and the estimation of the variance parameters 2.0%. Hence, a method that ignores the parameter estimation should not have a big effect on mean squared error estimation for this application. When we computed the PMSE using the asymptotic approximation and the bootstrap, we found that the estimates tracked each other closely for all HUCs, but the bootstrap estimates tended to be larger. Averaging the square root of the PMSE estimates over the HUCs, the difference between the bootstrap and the asymptotic estimates was approximately 8.5%. Hence, using the bootstrap method leads to potentially slightly conservative but (in our opinion) practically acceptable inference for this model.

In order to evaluate the appropriateness of the nonparametric bootstrap inference approach for this dataset more fully and compare it with alternative approaches, we performed a limited simulation study. Data were generated by taking the same covariates and spline basis functions as in this dataset, with parameter values set at those shown in Table 1 and with random components following known distributions. Two such distributions were investigated: (1) all random components were taken to be independent homoskedastic Gaussian random variables, and (2) they were generated as centered (zero-mean) and rescaled independent χ_1^2 random variables. For both cases, the true values of the PMSE of the nonparametric small area predictors were approximated by averaging the squared deviations between the true and predicted values for the small area means over 1000 realizations of the model. We considered four inference methods: the PMSE estimator from Theorem 3.1, a parametric bootstrap which uses normal distributions with REML-estimated variance parameters to generate the bootstrap random components, and two versions of the nonparametric bootstrap from Section 3.3, where we implemented the method both with and without setting $\mathbf{H} = \mathbf{0}$. All these inference methods were applied to a single realization of the model, with 1000 replicates for the bootstrap methods. For both the normally distributed and the χ_1^2 case, all methods produced estimated PMSE values that closely track the true values. In plots of estimated vs. true PMSE values across the HUCs (not shown here for brevity), all four methods displayed patterns that were very close to 45-degree lines. In addition to this visual assessment, we computed the squared deviations between the true root PMSE and each of the

estimators averaged those over the HUCs, and took the square root again. Finally, we divided these by the mean of the true root PMSEs to obtain relative measures of precision. In the normal case, these relative root deviations were 3.7% for the asymptotic method, 3.4% for the parametric bootstrap and 4.9% and 5.1% for the nonparametric bootstrap with and without inclusion of \mathbf{H} , respectively. For the situation with χ_1^2 distributions, the equivalent results are 5.7% for the asymptotic method, 6.0% for the parametric bootstrap, and 5.8% and 5.7% for the two nonparametric bootstraps methods. It therefore appears that the parametric bootstrap and the asymptotic method are more precise when the true distributions are normal, but all methods behave similarly when the distributions are non-Gaussian, with the parametric bootstrap performing slightly worse in that case. However, all these relative root deviations are small, so we conclude that, at least for the two setups considered here, all these methods are able to produce reliable PMSE estimators.

We also considered the distribution of likelihood ratio statistics, and compared the distribution over 1000 realizations of the simulation model with those obtained from the parametric and nonparametric bootstrap (without \mathbf{H}) for one model realization, and to a χ^2 mixture as discussed in Section 3.2. Figure 5 displays P-P plots for testing the significance of the spline and HUC effects using the different distribution estimators, for the situation where the random effects and errors follow centered and scaled χ_1^2 distributions. We only display the lower 20% of the values, since these are the ones relevant for hypothesis testing. For testing the spline effect, we see that the asymptotic approximation is severely biased, leading to a test that would reject the null hypothesis too often. The parametric and nonparametric bootstrap distributions perform better, with the former slightly more conservative. For the HUC effect, all three estimated distributions appear to be reasonable approximations of the true distribution, with the parametric bootstrap deviating the most from the true distribution. The results for the normally distributed random components (not shown) are similar, with the asymptotic distribution again clearly inappropriate for the spline effect test, but the other estimated distributions all close to the true distribution.

Both the spline and the HUC random effects are highly statistically significant as measured by the bootstrap-based likelihood ratio test, and the simulation results appear to indicate that the testing procedure is appropriate in this case. However, 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, especially since both capture spatial relationships between the observations. Table 2 shows

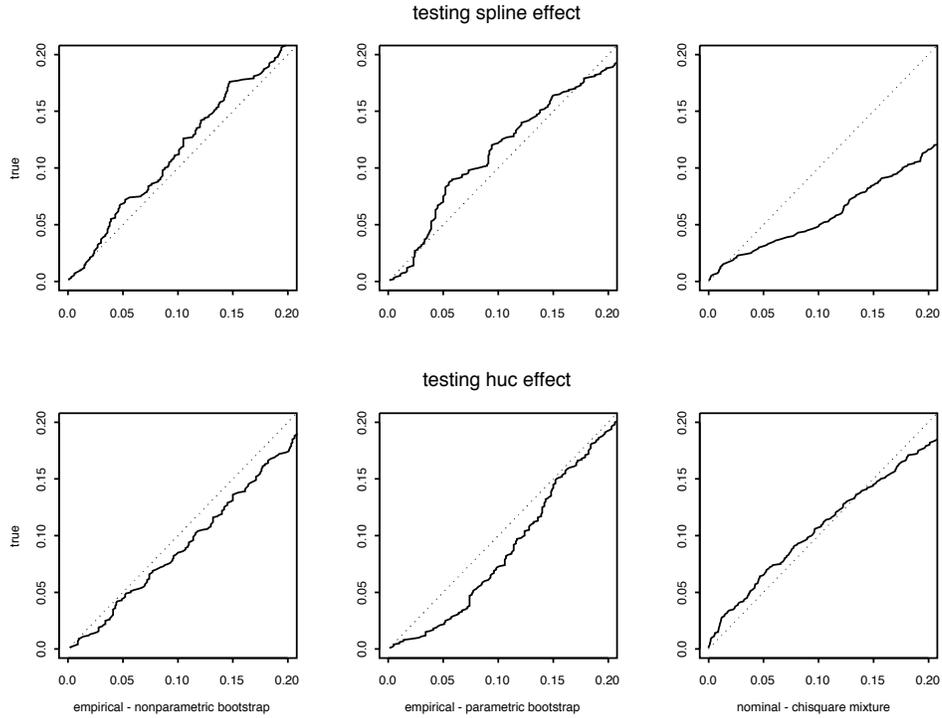


Figure 5: Partial P-P plots comparing the nonparametric bootstrap (left), parametric bootstrap (middle) and asymptotic χ^2 mixture (right) distribution of the likelihood ratio statistic with its simulated true distribution. Top row is for testing the spline effect, bottom row for the HUC effect.

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

Table 2: Comparison of AIC values and correlations between HUC model predictions and averages of the sample observations in the HUCs, for inclusion and exclusion of random effect terms in model; left number in each cell is AIC, right number is correlation.

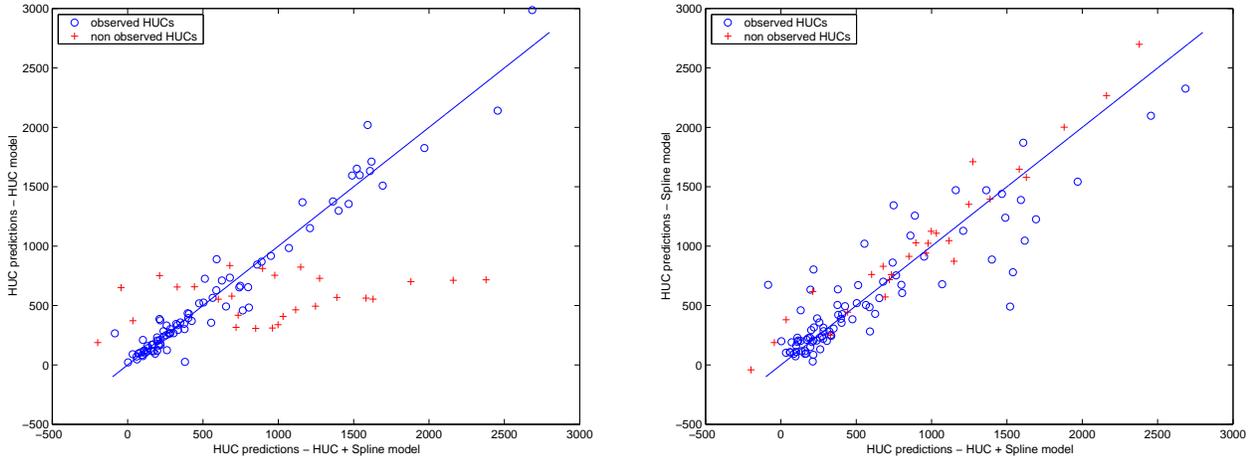


Figure 6: Comparison of HUC predictions for model with both random effects and models with single random effect (solid line is 45-degree line). Left plot: HUC only model; right plot: spline only model.

the correlations between the EBLUP \hat{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 with or without the addition of the spline random effect, while the smallest AIC is attained by the model with both random effects. 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. All three models with at least one random effect widely 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 Figure 6, we plot the HUC predictions obtained by the full model against those for the models with single random effects for a further comparison.

The plot on the left of Figure 6 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 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 2.

In the plot on the right of Figure 6, differences between the spline-only model and that with both random effects are not as obvious, but some large deviations from the 45-degree line are still 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 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 the plot, 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. The authors thank Marc Rogers for his assistance in producing the maps, and the Editor, Associate Editor and Referees for their helpful comments which resulted in substantial improvements to the manuscript.

Appendix A

Proof of Theorem 3.1.

This proof builds on the work of Das et al. (2004). First we compute the orders of magnitude of the quantities $d_j = \{\text{tr}(\mathbf{P}\mathbf{B}_j\mathbf{P}\mathbf{B}_j)\}^{1/2}$ ($j = 1, 2, 3$), which are, respectively, $O(\sqrt{n}/L_n)$, $O(n/L_n)$ and $O(T/L_n)$. The assumptions on L_n and T yield that the slowest rate is obtained by $d_1 = O(\sqrt{n}/L_n)$. Let the matrix $\mathbf{\Delta} = \text{diag}(d_1, d_2, d_3)$. Hence, the matrix elements of $-\mathbf{\Delta}^{-1}\mathcal{I}\mathbf{\Delta}^{-1}$ are all negative and bounded away from

infinity. Writing the variance matrix \mathbf{V} in the form $\mathbf{A}\mathbf{\Lambda}\mathbf{A}'$, where $\mathbf{\Lambda}$ is the diagonal matrix containing the eigenvalues and the matrix \mathbf{A} is orthogonal, with the entries of $\mathbf{\Lambda}$ of order $O(L_n)$, it follows that, for δ as in the assumption of the theorem,

$$\mathbb{E} \left(\sup_{\|\boldsymbol{\sigma}\| \leq L_n} |\hat{y}_t|^{2\delta} \right) = O(L_n^{-2\delta}) + O\{(T/n)^{2\delta}\}.$$

Let g be any number strictly larger than 8, then the above order is $O\{(\sqrt{n}/L_n)^{\delta_0}\}$, with $0 < \delta_0 < ((g-8)\delta - g)/4$. Next we compute the first and second partial derivatives of \hat{y}_t , with respect to $\boldsymbol{\sigma}^2$. Explicit expressions are easily obtained using the matrix formula as in (8). The vectors \mathbf{c}_t , $\hat{\boldsymbol{\beta}}$ and the matrices \mathbf{V} , $\boldsymbol{\Sigma}_w$ all depend on $\boldsymbol{\sigma}^2$. Straightforward computations yield that

$$\mathbb{E} \left(\left| \frac{\partial \hat{y}_t}{\partial \boldsymbol{\sigma}^2} \right|^{2g/(g-2)} \right) = O(L_n^{-2g/(g-2)}).$$

Take $\delta_1 = bg/(g-2)$ with $0 < b < 1/2$, then $\mathbb{E}(|\frac{\partial \hat{y}_t}{\partial \boldsymbol{\sigma}^2}|^{2g/(g-2)}) = O\{(\sqrt{n}/L_n)^{\delta_1}\}$. Similarly, we find that

$$\mathbb{E} \left(\sup_{\|\boldsymbol{\sigma}\| \leq L_n} \left\| \frac{\partial^2 \hat{y}_t}{\partial (\boldsymbol{\sigma}^2)^2} \right\|^2 \right) = O(L_n^{-2}) = O\{(\sqrt{n}/L_n)^{\delta_2}\},$$

with $0 < \delta_2 < 1/2$. Taken all this together is sufficient to obtain that

$$\mathbb{E}\{(\tilde{y}_t - \hat{y}_t)^2\} = \mathbb{E} \left\{ \left(\frac{\partial \hat{y}_t}{\partial \boldsymbol{\sigma}^2} \right)' \mathcal{I}^{-1} \frac{\partial \ell}{\partial \boldsymbol{\sigma}^2} \right\}^2 + o(L_n^2/n) = \text{tr}(\mathcal{S}'\mathbf{V}\mathcal{S}\mathcal{I}^{-1}) + o(L_n^2/n).$$

where ℓ denotes the REML log likelihood function, and the derivatives are all evaluated at the true parameter values (see Theorem 3.1 of Das et al. 2004). This proves the first part of the theorem. The second part of the theorem follows similarly as in Section 4 of Das et al. (2004).

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.
- Butar, F. B. and P. Lahiri (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference* 112, 63–76.

- Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statistics* 25, 573–578.
- 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.
- Clayton, D. and J. Kaldor (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43, 671–681.
- 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.
- Crainiceanu, C., D. Ruppert, G. Claeskens, and M. P. Wand (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* 92(1), 91–103.
- Das, K., J. Jiang, and J. N. K. Rao (2004). Mean squared error of empirical predictor. *Annals of Statistics* 32(2), 818–840.
- Datta, G. S. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* 10(2), 613–627.
- Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2), 89–121.
- Fay, Robert E., I. and R. A. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74, 269–277.
- Ghosh, M., K. Natarajan, T. W. F. Stroud, and B. P. Carlin (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*. 93, 273–282.
- Ghosh, M. and J. Rao (1994). Small area estimation: an appraisal. *Statistical Science* 9, 55–93.
- Hall, P. and T. Maiti (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, Series B* 62, 221–238.

- Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation. *Test* 15, 1–96.
- Kackar, R. N. and D. A. Harville (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* 79, 853–862.
- Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* 18, 199–210.
- 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.
- 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’Connell, 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.
- Pfeffermann, D. and H. Glickman (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. In *Proceedings of the Section on Survey Research Methods*, Washington, DC, pp. 4167–4178. American Statistical Association.
- Prasad, N. G. N. and J. N. K. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* 85, 163–171.
- Rao, J. N. K. (2003). *Small Area Estimation*. New York: Wiley.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 11, 735–757.

- Ruppert, R., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.
- Self, S. G. and K. Y. Liang (1987). Asymptotic properties of maximum likelihood and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*. 82, 605–610.
- Sen, P. K. and M. J. Silvapulle (2002). An appraisal of some aspects of statistical inference under inequality constraints. *J. Statist. Plann. Inference* 107(1-2), 3–43.
- Stram, D. O. and J. W. Lee (1994). Variance component testing in the longitudinal mixed effects model. *Biometrics* 50, 1171–1177.
- Vu, H. T. V. and S. Zhou (1997). Generalization of likelihood ratio tests under nonstandard conditions. *Annals of Statistics* 25, 897–916.
- 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 (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology* 30, 209–218.