Survey Methodology, June 2007
Vol. 33, No. 1, pp. 35-44
Statistics Canada, Catalogue No. 12-001

Semiparametric model-assisted estimation for natural resource surveys

F. Jay Breidt, Jean D. Opsomer, Alicia A. Johnson and M. Giovanna Ranalli [1]

Abstract

Auxiliary information is often used to improve the precision of survey estimators of finite population means and totals through ratio or linear regression estimation techniques. Resulting estimators have good theoretical and practical properties, including invariance, calibration and design consistency. However, it is not always clear that ratio or linear models are good approximations to the true relationship between the auxiliary variables and the variable of interest in the survey, resulting in efficiency loss when the model is not appropriate. In this article, we explain how regression estimation can be extended to incorporate semiparametric regression models, in both simple and more complicated designs. While maintaining the good theoretical and practical properties of the linear models, semiparametric models are better able to capture complicated relationships between variables. This often results in substantial gains in efficiency. The applicability of the approach for complex designs using multiple types of auxiliary variables will be illustrated by estimating several acidification-related characteristics for a survey of lakes in the Northeastern US.

Key Words: Regression estimation; Smoothing; Kernel regression; Lake chemistry.

1. Introduction

Post-stratification, calibration and regression estimation are different design-based approaches that can be used to improve the precision of estimators when auxiliary information is available at the estimation stage. *Model-assisted estimation* (Särndal, Swensson and Wretman 1992) provides a convenient framework in which to develop these and related survey estimators. Under that framework, a superpopulation model describes the relationship between the variable of interest and the auxiliary variables. This model is then used to construct sample-based estimators that have improved precision when the model is correct, but maintain key design properties such as consistency and an estimable variance when the model is incorrect.

Until recently, the superpopulation models used in this context were formulated as parametric models, most often ratio or linear models. While reasonable in many practical applications, there are also many situations in which such relatively simple models are not good representations of the relationship between the variable of interest and the auxiliary variables. In Breidt and Opsomer (2000), a nonparametric model-assisted estimator was proposed based on local polynomial regression, which generalized the well-established parametric regression estimators. With this estimator, the superpopulation is no longer required to follow a pre-specified parametric shape. Instead, the relationship between the the variable(s) of interest in the survey and the auxiliary variable is required to be smooth (continuous), but is otherwise left completely unspecified.

In the current paper, we formally extend the theory of Breidt and Opsomer (2000) to the semiparametric regression context, in which some variables are incorporated linearly, and others are incorporated through smooth additive terms. This extension makes their results more useful in practice, since auxiliary information is very often multidimensional in nature, and almost always contains categoryical variables that need to enter the regression model parametrically (through the use of indicator variables). An illustration of this is provided by a survey of lakes in the Northeastern states of the U.S. conducted by the Environmental Monitoring and Assessment Program of the US Environmental Protection Agency. In that survey, 334 lakes were sampled from a population of 21,026 lakes between 1991 and 1996. We will apply the semiparametric model-assisted estimator to produce estimates of the mean and distribution function of the *acid neutralizing capacity* and other chemistry variables of interest. In this application, we will include in the model both categorical and continuous variables linearly and a continuous variable as a smooth additive term.

In Opsomer, Breidt, Moisen and Kauermann (2007), the nonparametric model-assisted estimation principle was extended to generalized additive models (GAMs) and applied in an interaction model for the estimation of variables from Forest Inventory and Analysis surveys. While GAMs also contained a mixture of categorical (parametric) and nonparametric terms, a complete theoretical development is not possible in the case of GAMs, and was therefore not provided there. The semiparametric model considered in this article can be viewed as a special case of a GAM with an identity link function. Unlike the

[1] F. Jay Breidt, Department of Statistics, Colorado State University, Fort Collins CO 80523, U.S.A.; Jean D. Opsomer, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. E-Mail: jopsomer@iastate.edu; Alicia A. Johnson, School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis MN 55455, U.S.A.; M. Giovanna Ranalli, Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Pascoli, 06123 Perugia, Italy.

"general" GAM, the semiparametric model allows for formal derivation of the statistical properties of the model-assisted estimator.

The remainder of the article is structured as follows. In Section 2, the semiparametric model-assisted estimator is defined. Section 3 states and proves the design properties of the estimator. Section 4 describes the application of semiparametric model-assisted estimation to the Northeastern Lakes data. Section 5 provides a conclusion.

2. Semiparametric model-assisted estimation

We begin by considering the superpopulation model with a single univariate nonparametric term and a parametric component; extension to several nonparametric terms is addressed in Section 3.2. The parametric component can be composed of an arbitrary number of linear terms. This model is the semiparametric model studied by Speckman (1988), among others. This superpopulation model, which we denote by ξ, can be written down as

$$\mathrm{E}_\xi(y_k) = g(x_k,\ z_k) = m(x_k) + z_k\boldsymbol{\beta}$$
$$\mathrm{Var}_\xi(y_k) = v(x_k,\ z_k) \qquad (1)$$

with x_k a continuous auxiliary variable to be modelled nonparametrically and $z_k = (z_{1k}, ..., z_{Dk})$ a vector of D categorical or continuous auxiliary variables that are parametrically specified. The functions $m(\cdot)$ and $v(\cdot,\cdot)$ and the parameter vector $\boldsymbol{\beta}$ are unknown. For identifiability purposes, we will assume that the vector z_k contains an intercept term, and that the function $m(\cdot)$ is centered around 0 with respect to the distribution of the x_k. We will derive the model-assisted estimator that uses model (1) by first defining population-level estimators for the unknown functions and parameters, and then constructing sample-based estimators. This is the same approach used for the parametric case in Särndal *et al.* (1992, Chapter 6).

Let $U = \{1, 2, ..., N\}$ represent the ordered labels for a finite population of interest. As the population estimator for $g(x_k, z_k)$, we will use the *backfitting estimator* described in Opsomer and Ruppert (1999). We first introduce the required notation. Let $K(\cdot)$ represent a kernel function used to define the neighborhoods in which the local polynomials will be fitted (assumptions on K are specified in the Appendix). The population *smoother vector* for local polynomial regression of degree p at x_k is defined as

$$\boldsymbol{s}_{Uk}^T = \boldsymbol{e}_1^T(\boldsymbol{X}_{Uk}^T\boldsymbol{W}_{Uk}\boldsymbol{X}_{Uk})^{-1}\boldsymbol{X}_{Uk}^T\boldsymbol{W}_{Uk}$$

with $\boldsymbol{e}_1$ a vector of length $p+1$ with a 1 in the first position and 0s elsewhere, $\boldsymbol{W}_{Uk} = \mathrm{diag}\{h^{-1}K((x_1 - x_k)/h), ..., h^{-1}K((x_N - x_k)/h)\}$ and

$$\boldsymbol{X}_{Uk} = \begin{bmatrix} 1 & x_1 - x_k & \dots & (x_1 - x_k)^p \\ \vdots & & \ddots & \vdots \\ 1 & x_N - x_k & \dots & (x_N - x_k)^p \end{bmatrix}.$$

The smoother $\boldsymbol{s}_{Uk}$ can be applied to the vector $\boldsymbol{Y}_U = (y_1, ..., y_N)^T$ to produce the nonparametric regression fit with respect to the variable x at observation x_k. It can also be applied to any of the columns of $\boldsymbol{Z}_U = (z_1^T, ..., z_N^T)^T$ to smooth those with respect to x. This will be done in the derivation of the properties of the semiparametric estimator (Section 3).

In addition to the smoother vector at x_k, $\boldsymbol{s}_{Uk}^T$, we also need to define the *smoother matrix* at all the observation points $x_1, ..., x_N$,

$$\boldsymbol{S}_U = \begin{bmatrix} \boldsymbol{s}_{U1}^T \\ \vdots \\ \boldsymbol{s}_{UN}^T \end{bmatrix},$$

and the *centered smoother matrix* $\boldsymbol{S}_U^* = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/N)\boldsymbol{S}_U$. When the smoother matrix is applied to $\boldsymbol{Y}_U$, it produces the vector of nonparametric regression fits at all the observation points. The centered smoother matrix $\boldsymbol{S}_U^*$ produces centered fits, *i.e.*, the overall mean of the fitted values is subtracted from each fitted value. The centering is used to maintain identifiability of the estimators, as explained in Opsomer and Ruppert (1999).

For any observation x_k, a possible estimator of $m(x_k)$ could be defined as $\boldsymbol{s}_{Uk}^T\boldsymbol{Y}_U$, with or without a centering adjustment. This estimator would generally be poor, since it does not take into account the fact that the y_k contain a parametric component that depends on the z_k. A more efficient estimator is provided by jointly estimating both $m(\cdot)$ and $\boldsymbol{\beta}$, as is done by the following set of estimators

$$\boldsymbol{B} = (\boldsymbol{Z}_U^T(\boldsymbol{I} - \boldsymbol{S}_U^*)\boldsymbol{Z}_U)^{-1}\boldsymbol{Z}_U^T(\boldsymbol{I} - \boldsymbol{S}_U^*)\boldsymbol{Y}_U$$
$$m_k = \boldsymbol{s}_{Uk}^T(\boldsymbol{Y}_U - \boldsymbol{Z}_U\boldsymbol{B}) \quad k = 1, ..., N. \qquad (2)$$

In these estimators, $\boldsymbol{B}$ is calculated first, and then the "residual vector" $\boldsymbol{Y}_U - \boldsymbol{Z}_U\boldsymbol{B}$ is smoothed with respect to x. The estimators in (2) are identical to the *backfitting estimators* for additive models described in Hastie and Tibshirani (1990) and implemented in `gam` in S-Plus, R or SAS. As a population estimator for $\mathrm{E}_\xi(y_k) = g(x \mid k, z_k)$, we use

$$g_k = m_k + z_k\boldsymbol{B}.$$

We now explain how to construct a model-assisted estimator based on the semiparametric regression approach. Let $A \subset U$ be a sample of size n drawn from U according to sampling design $p(A)$ with one-way and two-way inclusion probabilities $\pi_k = \Sigma_{A \ni k}\, p(A)$, $\pi_{kl} = \Sigma_{A \ni k,l}\, p(A)$, respectively. If the g_k, $k = 1, \ldots, N$ were available, it would be possible to construct a *difference estimator* for the population mean of the y_k, $\bar{y}_N = \Sigma_U\, y_k / N$, as

$$\hat{y}_{\text{dif}} = \frac{1}{N}\sum_U g_k + \frac{1}{N}\sum_A \frac{y_k - g_k}{\pi_k}, \qquad (3)$$

which is design unbiased and has design variance

$$\text{Var}_p(\hat{y}_{\text{dif}}) = \frac{1}{N^2}\sum\sum_U (\pi_{kl} - \pi_k \pi_l)\frac{y_k - g_k}{\pi_k}\,\frac{y_l - g_l}{\pi_l}$$

(Särndal *et al.* 1992, page 221). The design variance is small if the deviations between y_k and g_k are small. This estimator is not feasible, since it requires knowledge of all the x_k, $\boldsymbol{z}_k$ and y_k for the population to calculate. Instead, we will construct a feasible estimator by replacing the g_k by sample-based estimators. The sample-based estimators corresponding to the population estimators in (2) are constructed as follows. The design-weighted local polynomial smoother vector is

$$\boldsymbol{s}_{Ak}^{0T} = \boldsymbol{e}_1^T (\boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak} \boldsymbol{X}_{Ak})^{-1} \boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak}, \qquad (4)$$

with $\boldsymbol{X}_{Ak}$ containing the rows of $\boldsymbol{X}_{Uk}$ corresponding to the $k \in A$ and

$$\boldsymbol{W}_{Ak} = \text{diag}\left\{\frac{1}{\pi_j h} K\left(\frac{x_j - x_k}{h}\right) : j \in A\right\}.$$

The matrix $\boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak} \boldsymbol{X}_{Ak}$ in (4) will be singular if, for some sample A, there are less than $p+1$ observations in the support of the kernel at some x_k. This issue can be avoided in practice by selecting a bandwidth large enough to make that matrix invertible. However, this situation cannot be excluded in general and we need an estimator that exists for every sample A for the theoretical derivations of Section 3. Hence, we will consider the following adjusted sample smoother vector

$$\boldsymbol{s}_{Ak}^T = \boldsymbol{e}_1^T (\boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak} \boldsymbol{X}_{Ak} + \text{diag}(\delta N^{-2}))^{-1} \boldsymbol{X}_{Ak}^T \boldsymbol{W}_{Ak}, \qquad (5)$$

for some small $\delta > 0$, as done in Breidt and Opsomer (2000). The sample smoother matrix and its centered version are

$$\boldsymbol{S}_A = [\boldsymbol{s}_{Ak}^T : k \in A] \quad \boldsymbol{S}_A^* = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T \boldsymbol{\Pi}_A^{-1}/N)\boldsymbol{S}_A$$

with $\boldsymbol{\Pi}_A = \text{diag}\{\pi_k : k \in A\}$. The design-weighted estimators for $\boldsymbol{B}$ and the m_k are

$$\hat{\boldsymbol{B}} = (\boldsymbol{Z}_A^T \boldsymbol{\Pi}_A^{-1}(\boldsymbol{I} - \boldsymbol{S}_A^*)\boldsymbol{Z}_A)^{-1} \boldsymbol{Z}_A^T \boldsymbol{\Pi}_A^{-1}(\boldsymbol{I} - \boldsymbol{S}_A^*)\boldsymbol{Y}_A \qquad (6)$$

$$\hat{m}_k = \boldsymbol{s}_{Ak}^T(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T \hat{\boldsymbol{B}}), \qquad (7)$$

where $\boldsymbol{Z}_A$ and $\boldsymbol{Y}_A$ denote the sample versions of $\boldsymbol{Z}_U$ and $\boldsymbol{Y}_U$, respectively. Note that the estimator $\hat{m}_k$ is defined for any x_k in the population, not only those appearing in the sample. As for the population estimators, these estimators can again be written as the solution to backfitting equations, so that they can be calculated by appropriately weighted versions of the existing algorithms.The estimator for g_k is

$$\hat{g}_k = \hat{m}_k + \boldsymbol{z}_k \hat{\boldsymbol{B}}.$$

The semiparametric model-assisted estimator is then constructed by replacing the g_k in (3) by the $\hat{g}_k$:

$$\hat{y}_{\text{reg}} = \frac{1}{N}\sum_U \hat{g}_k + \frac{1}{N}\sum_A \frac{y_k - \hat{g}_k}{\pi_k}. \qquad (8)$$

Defining $\bar{y}_\pi = \Sigma_A\, y_k / \pi_k$ and similarly for $\bar{\boldsymbol{z}}_\pi$, an equivalent expression for $\hat{y}_{\text{reg}}$ is given by

$$\hat{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\boldsymbol{z}}_N - \bar{\boldsymbol{z}}_\pi)\hat{\boldsymbol{B}} + \frac{1}{N}\sum_U \hat{m}_k - \frac{1}{N}\sum_A \frac{\hat{m}_k}{\pi_k}, \quad (9)$$

which shows that the semiparametric estimator can be interpreted as a "traditional" linear regression survey estimator using the parametric model component $\boldsymbol{z}\boldsymbol{\beta}$, with an additional correction term for the nonparametric component of the model. This estimator also shares some desirable properties with the fully parametric regression estimators. It is location and scale invariant, and it is calibrated for both the parametric and the nonparametric model components, in the sense that $\hat{x}_{\text{reg}} = \bar{x}_N$ and $\hat{\boldsymbol{z}}_{\text{reg}} = \bar{\boldsymbol{z}}_N$. The calibration for the variables in the parametric term can be checked directly by using expressions (6) and (7), while the calibration for the nonparametrically specified variable x_k follows from the fact that $\boldsymbol{s}_{Ak}^T \boldsymbol{X}_A = x_k$, where $\boldsymbol{X}_A = (x_k: k \in A)^T$ (we are ignoring the effect of the adjustment $\text{diag}(\delta N^{-2})$ in (5), because that adjustment can be made arbitrarily small). In addition, the estimator can be written as a weighted sum of the y_k, $k \in A$, so that a set of weights w_k can be obtained and applied to any survey variable of interest.

3. Properties and extensions

3.1 Design properties

In this section, we explore the design properties of the semiparametric estimator (8). In particular, we prove that $\hat{y}_{\text{reg}}$ is design $\sqrt{n}$-consistent, and we derive its asymptotic distribution, including an estimated variance. This will be done in the design-asymptotic context used in Isaki and Fuller (1982) and in Breidt and Opsomer (2000), in which both the population and the samples increase in size as $N \to \infty$. All proofs and the necessary assumptions are in the Appendix.

In the following theorem, we prove the design consistency of the semiparametric estimator. We also show that the convergence rate is $\sqrt{n}$, the usual rate for design estimators.

Theorem 3.1 *Under the assumptions* $A1-A8$, *the estimator* $\hat{y}_{\text{reg}}$ *in* (8) *is design consistent with rate* $\sqrt{n}$, *in the sense that*

$$\hat{y}_{\text{reg}} = \bar{y}_N + O_p\left(\frac{1}{\sqrt{n}}\right).$$

The following theorem proves that a central limit theorem for $\hat{y}_{\text{reg}}$ exists whenever it exists for the expansion estimator $\bar{y}_\pi$.

Theorem 3.2 *Under the assumptions* $A1$ - $A8$, *if*

$$\frac{\bar{y}_\pi - \bar{y}_N}{\sqrt{\hat{V}(\bar{y}_\pi)}} \rightarrow N(0,\ 1),$$

with

$$\hat{V}(\bar{y}_\pi) = \frac{1}{N^2} \sum\sum_{A} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

for a given sampling design, then we also have

$$\frac{\hat{y}_{\text{reg}} - \bar{y}_N}{\sqrt{\hat{V}(\hat{y}_{\text{reg}})}} \rightarrow N(0,\ 1),$$

with

$$\hat{V}(\hat{y}_{\text{reg}}) = \frac{1}{N^2} \sum\sum_{A} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k - \hat{g}_k}{\pi_k} \frac{y_l - \hat{g}_l}{\pi_l}. \quad (10)$$

3.2 Semiparametric additive model

The results in Theorems 3.1 and 3.2 use the semiparametric model (1), which contains a single univariate nonparametric term $m(\cdot)$. In many practical applications, several auxiliary variables will be available that could be included in the nonparametric portion of a model, but the curse of dimensionality makes it often difficult to combine several variables into a single multi-dimensional nonparametric term. Instead, the variables that are to be included nonparametrically will be treated as univariate components. This results in the *semiparametric additive model*, which is written as

$$\begin{aligned} \mathrm{E}_\xi(y_k) &= g(\boldsymbol{x}_k,\ z_k) = m_1(x_{1k}) + \ldots + m_Q(x_{Qk}) + z_k\,\boldsymbol{\beta} \\ \mathrm{Var}_\xi(y_k) &= v(\boldsymbol{x}_k,\ z_k) \end{aligned}$$

where the $m_q(\cdot)$, $q = 1, \ldots, Q$ and $v(\cdot, \cdot)$ are unknown smooth functions.

When $Q = 2$, expressions similar to (6) and (7) can be developed, using the additive model decompositions of Opsomer and Ruppert (1997), and for $Q > 2$, recursive expressions can be derived using the approach of Opsomer (2000). The estimator would then be written as in equations (6) and (7), but with the smoother vectors $\boldsymbol{s}_{Ak}$ and smoother matrix $\boldsymbol{S}_A$ replaced by complicated higher-dimensional additive model smoothers (see Opsomer (2000) for details). Because of this, formally proving the properties of the model-assisted estimator for the case with arbitrary Q would be a challenging task beyond the scope of the current article.

In practice, the backfitting algorithm formulation provides a much more efficient and simple way to calculate the semiparametric estimator. Let $\boldsymbol{s}_{Aqk}$ represent the sample smoother vector, as defined in (5), for the variable x_q at the observation x_{qk} and $\boldsymbol{S}_{Aq}$ is the corresponding smoother matrix for the variable x_q. Also, $\hat{m}_{qk}$ denotes the sample-weighted backfitting estimator for $m_q(x_{qk})$ and $\hat{\boldsymbol{m}}_{Aq} = (\hat{m}_{qk},\ k \in A)$. The backfitting algorithm for a model including Q nonparametric terms consists of the following set of equations, iterated to converge:

$$\begin{aligned} \hat{\boldsymbol{B}} &= (\boldsymbol{Z}_A^T \boldsymbol{\Pi}_A^{-1} \boldsymbol{Z}_A)^{-1} \boldsymbol{Z}_A^T \boldsymbol{\Pi}_A^{-1} \left(\boldsymbol{Y}_A - \sum_{q=1}^{Q} \hat{\boldsymbol{m}}_{Aq} \right) \\ \hat{\boldsymbol{m}}_{A1} &= \boldsymbol{S}_{A1} \left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T \hat{\boldsymbol{B}} - \sum_{q \neq 1} \hat{\boldsymbol{m}}_{Aq} \right) \\ \vdots \quad & \quad \vdots \\ \hat{\boldsymbol{m}}_{AQ} &= \boldsymbol{S}_{AQ} \left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T \hat{\boldsymbol{B}} - \sum_{q \neq Q} \hat{\boldsymbol{m}}_{Aq} \right). \end{aligned}$$

These equations provide weighted fits at the sample locations $k \in A$ only. For the remaining locations $k \in U$ not in A, an additional smoothing step is required after obtaining the $\hat{\boldsymbol{m}}_{Aq}$, $q = 1, \ldots, Q$:

$$\hat{m}_{kq} = \boldsymbol{s}_{Aqk}^T \left(\boldsymbol{Y}_A - \boldsymbol{Z}_A^T \hat{\boldsymbol{B}} - \sum_{q' \neq q} \hat{\boldsymbol{m}}_{Aq'} \right).$$

The sample-based estimators for the mean function at all $k \in U$ are then defined as $\hat{g}_k = \hat{m}_{k1} + \ldots + \hat{m}_{kQ} + z_k \hat{\boldsymbol{B}}$, which are used in expression (8) to construct the model-assisted estimator.

4. Application to Northeastern Lakes survey

In this section, we will show the applicability of the semiparametric regression estimator on a dataset of water chemistry samples. As will be illustrated, once a set of auxiliary variables and a model has been selected, computing survey estimators for the semiparametric model is as easy as for linear models, and hence can lead to improved precision for relatively little cost.

The National Surface Water Survey (NSWS) sponsored by the U.S. Environmental Protection Agency (EPA) between the years of 1984 and 1986 estimated 4.2 percent of the lakes in the northeastern region of the United States to be acidic (Stoddard, Kahl, Deviney, DeWalle, Driscoll, Herlihy, Kellogg, Murdoch, Webb and Webster 2003). Acid-sensitive Northeastern lakes were among the concerns addressed by the Clean Air Act Amendment (CAAA) of 1990, which placed restrictions on industrial sulfur and nitrogen emissions in an effort to reduce the acidity of these waters. A common measurement of acidity is acid neutralizing capacity (ANC), which is defined as a water's ability to buffer acid. An ANC value less than zero $\mu\text{eq}/L$ indicates that the water has lost all ability to buffer acid. Surface waters with ANC values below 200 $\mu\text{eq}/L$ are considered at risk of acidification, and values less than 50 $\mu\text{eq}/L$ are considered at high risk (National Acid Precipitation Assessment Program (1991), page 15).

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. These data were collected in order to determine the effect that restrictions put in place by the CAAA had on the ecological condition of these waters. 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. Multiple measurements on the same lake were averaged in order to obtain one measurement per lake sampled. Lakes to be included in the survey were selected using a complex sampling design commonly employed by EMAP based on a hexagonal grid frame (see Larsen, Thornton, Urquhart and Paulsen (1993) for a description of the sampling design).

Let y_k represent the (possibly averaged) ANC value of the k^{th} sampled lake. A very simple estimate of the ANC mean of the lakes is represented by the expansion estimator $\bar{y}_\pi$. In this as in many surveys, a better choice is the Hájek estimator,

$$\hat{y}_H = \frac{1}{\hat{N}} \sum_{k \in A} \frac{y_k}{\pi_k}, \tag{11}$$

which applies a ratio type adjustment for the estimation of the population size through $\hat{N} = \sum_{k \in A} 1/\pi_k$. However, auxiliary variables are available for each lake in this population, so that it should be possible to further improve upon the efficiency of the Hájek estimator. The following variables are available for each $k \in U$:

x_k = UTMX, x-geographical coordinate of the centroid of each lake in the UTM coordinate system,

$z_{j,k}$ = indicator variable for eco-region $j = 1, \ldots, 6$,

$z_{7,k}$ = UTMY, y-geographical coordinate,

$z_{8,k}$ = elevation.

There are seven different eco-regions included in the population, thus dummy variables $z_{j,k}$ are constructed for $j = 1, \ldots, 6$. A semiparametric regression estimator for the variable y will be constructed by treating the UTMX variable x as a nonparametric term and the remaining variables $z_1 - z_8$ as a parametric component. Model selection was used to determine that treating the other two continuous variables as nonparametric did not improve the model fit. For comparison purposes, we also computed a regression estimator that treats all terms as parametric. This estimator is therefore identical to the semiparametric estimator, except that the x-geographical coordinate is modeled linearly. We will denote this fully parametric regression estimator by $\hat{y}_{\text{par}}$.

In order to determine the estimated efficiency of survey estimators, we need to compute the variance estimates. However, second order inclusion probabilities were not available, thus we cannot evaluate $\hat{V}(\hat{y}_{\text{reg}})$ as in (10). In order to come up with appropriate variance estimates, we treat the complex sampling design as a stratified sample taken with replacement. The 14 strata we selected correspond to groups of spatial clusters of lakes that appeared in the original design, and that were used to ensure spatial distribution of the sampled lakes over the region of interest. Larsen *et al.* (1993) provide details on the construction of the spatial clusters.

Let H be the number of strata, n_h the number of observations within stratum h, and A_h the set of sampled elements that fall in stratum h. Define $p_k = n_h^{-1} \pi_k$. Using this notation and the assumption of a stratified sample with replacement, we rewrite the semiparametric estimator as

$$\begin{aligned} \hat{y}_{\text{reg}} &= \frac{1}{N} \sum_{k \in U} \hat{g}(x_k, z_k) \\ &\quad + \frac{1}{N} \sum_{h \in H} \frac{1}{n_h} \sum_{k \in A_h} \frac{y_k - \hat{g}(x_k, z_k)}{p_k} \end{aligned} \tag{12}$$

and the variance estimator as

$$\hat{V}(\hat{y}_{\text{reg}}) = \frac{1}{N^2} \sum_{h \in H} S_h^2,$$

where S_h^2 is the estimated within-stratum weighted residual variance for stratum h. Assuming the strata are sampled with replacement, Särndal *et al.* (1992, page 421 - 422) suggest S_h^2 can be calculated as

$$S_h^2 = \frac{1}{n_h(n_h - 1)} \sum_{k \in A_h} \left(\begin{array}{c} \dfrac{y_k - \hat{g}(x_k, z_k)}{p_k} \\ - \displaystyle\sum_{l \in A_h} \dfrac{y_l - \hat{g}(x_l, z_l)}{\pi_l} \end{array} \right)^2 . \tag{13}$$

Similarly, we estimate $\hat V(\hat y_H)$ through

$$\hat V(\hat y_H) = \frac{1}{\hat N^2}\sum_{h\in H}\frac{1}{n_h(n_h-1)}\sum_{k\in A_h}\left(\frac{y_k-\hat y_H}{p_k} - \sum_{l\in A_h}\frac{y_l-\hat y_H}{\pi_l}\right)^2, \qquad (14)$$

and the expression for $\hat V(\hat y_{\text{par}})$ is obtained completely analogously as for $\hat V(\hat y_{\text{reg}})$ except that $\hat g(x_k, z_k)$ is computed by linear regression.

This setup allows us to obtain the following estimates of mean ANC for the Northeastern lakes, together with variance estimates and approximate 95% confidence intervals (CI). A local linear fit has been employed for the nonparametric term with bandwidth set at one tenth of the range of UTMX.

$$\hat y_{\text{reg}} = 558.0\ \mu\text{eq}/\text{L} \quad \hat V(\hat y_{\text{reg}}) = 2534.6 \quad \text{CI} = (459.3; 656.6)$$

$$\hat y_{\text{par}} = 577.3\ \mu\text{eq}/\text{L} \quad \hat V(\hat y_{\text{par}}) = 3239.6 \quad \text{CI} = (465.8; 688.9)$$

$$\hat y_H = 555.9\ \mu\text{eq}/\text{L} \quad \hat V(\hat y_H) = 4313.3 \quad \text{CI} = (427.2; 684.7)$$

The confidence interval constructed using the Hájek estimator is about 31% wider than that constructed using the semiparametric estimator, while the interval for the fully parametric regression estimator is 13% wider. These results show evidence of an improvement in efficiency provided by accounting for the auxiliary information in both a parametric and nonparametric way in the mean estimation procedure, with the nonparametric estimator able to capture some additional efficiency beyond that of the parametric estimator.

As mentioned above, an important goal of this application is the assessment of how many lakes are at risk of acidification or are acidified already. That is, we are interested in estimating the proportion of Northeastern lakes with ANC values smaller than some specific threshold values. We can determine such proportions by estimating the finite population distribution function,

$$F_N(t) = \frac{1}{N}\sum_{k\in U} I_{\{y_k\le t\}}$$

at specific threshold values t, where $I_{\{y_k\le t\}}$ denotes the indicator function taking a value of 1 if $y_k \le t$ and 0 otherwise. Because all three estimators can be expressed as weighted sums of sample observations, the weights obtained for each can be applied directly to the $I_{\{y_k\le t\}}$ for the sample to estimate $F_N(t)$ for any desired t. Let us denote by $\hat F_H(t)$, $\hat F_{\text{reg}}(t)$ and $\hat F_{\text{par}}(t)$ the Hájek, semiparametric and parametric regression estimators of the distribution function, respectively. Estimates for their design variances are computed by plugging the indicator variables in equations (13) and (14).

Figure 1 shows estimates of the ANC cdf produced by $\hat F_H(t)$, $\hat F_{\text{par}}(t)$ and $\hat F_{\text{reg}}(t)$ evaluated on a grid of 1,000 equally spaced values for t. Included are their respective pointwise 95% confidence intervals calculated at each grid point. All three estimators are similar, but the confidence bands for the parametric and semiparametric regression estimators tend to be narrower. Averaged over all 1,000 grid points, the widths of the confidence bands are 0.093 for $\hat F_H(t)$, 0.084 for $\hat F_{\text{par}}(t)$ and 0.075 for $\hat F_{\text{reg}}(t)$, respectively.

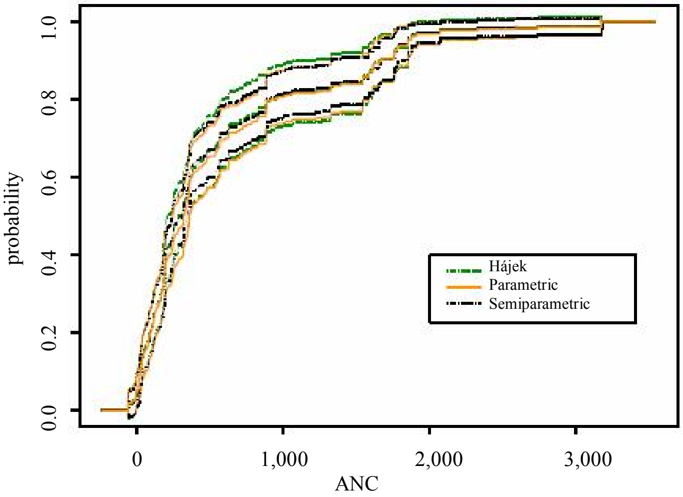

Figure 1
Estimates of the population cumulative distribution function for ANC and confidence bounds produced by Hájek, parametric and semiparametric regression estimators

Along with ANC, the EMAP survey of Northeastern lakes measured the concentration of multiple chemistry variables including sulfate, magnesium and chloride, so that the survey weights obtained for ANC can also be applied to these concentrations as well as their respective cdfs. As another illustration of the semiparametric estimation approach, it is possible to "invert" $\hat F_{\text{reg}}(t)$ to obtain quantile estimators $\hat\theta_{\text{reg}}(\alpha) = \min\{t : \hat F_{\text{reg}}(t) \ge \alpha\}$ of these additional chemistry variables. Table 1 displays semiparametric estimates of the first, second, and third quartiles of sulfate, magnesium, and chloride measured in $(\mu\text{eq}/L)$. Variance estimation for these quantiles could be handled using asymptotic results of Francisco and Fuller (1991), but will not be explored further here.

Table 1 Quartile estimates of chemistry variables

α	Sulfate	Magnesium	Chloride
0.25	73.3	63.8	27.4
0.50	104.3	127.0	162.2
0.75	201.4	221.9	462.2

5. Conclusion

In this article, we have described a model-assisted estimator that uses semiparametric regression to capture relationships between multiple population-level auxiliary variables and the survey variables. We have developed asymptotic theory that shows the resulting estimator is design consistent and asymptotically normal under mild conditions on the design and the population. This generalizes the results of Breidt and Opsomer (2000), who had proved similar results for a univariate nonparametric model-assisted estimator. The semiparametric estimator was applied to data from a survey of lakes in the Northeastern U.S., where it was shown to be more efficient than an estimator that does not take advantage of the auxiliary variables and than a fully parametric regression estimator.

In addition to its theoretical properties, the semiparametric model-assisted estimator has attractive practical properties as well. As noted earlier, it is fully calibrated for the auxiliary variables, whether used in the parametric or nonparametric model components, and it is location and scale invariant. The estimator can be expressed as a weighted sum of the sample observations, so that it conforms to the traditional survey estimation paradigm and a single set of weights can be applied to all the survey variables, hence preserving relationships between variables.

One issue which was not addressed in the current article is the selection of the smoothing parameter for the nonparametric component of the regression model. This is a challenging topic in the model-assisted context, further complicated by the just mentioned fact that a single set of survey regression weights is applied to all the survey variables: because the optimal bandwidth choice depends on the variable being smoothed, no single bandwidth (and hence set of weights) will be optimal for all variables in the survey. This topic is currently being explored by the authors.

Acknowledgments

The research for this article was supported by National Science Foundation grants DMS-0204531 and DMS-0204642, and by STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon State University. This manuscript 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.

Appendix

Technical assumptions and derivations

We begin by stating the necessary assumptions, which extend those used in Breidt and Opsomer (2000) to the semiparametric model.

Assumptions:

- **A1** *Distribution of the errors under* ξ: *the errors* ε_k *are independent and have mean zero*, *variance* $v(x_k, z_k)$, *and compact support, uniformly for all* N.
- **A2** *Distribution of the covariates*: *the* x_k *and* z_k *are considered fixed with respect to the superpopulation model* ξ. *The* z_k *are assumed to have bounded support, and the* x_k *are independent and identically distributed* $F(x) = \int_{-\infty}^{x} f(t)\,dt$, *where* $f(\cdot)$ *is a density with compact support* $[a_x, b_x]$ *and* $f(x) > 0$ *for all* $x \in [a_x, b_x]$.
- **A3** *Nonparametric mean and variance functions*: *the mean function* $m(\cdot)$ *is continuous, and the variance function* $v(\cdot, \cdot)$ *is bounded and strictly greater than* 0.
- **A4** *Kernel* K: *the kernel* $K(\cdot)$ *has compact support* $[-1, 1]$, *is symmetric and continuous*, *and satisfies* $\int_{-1}^{1} K(u)\,du = 1$.
- **A5** *Sampling rate* nN^{-1} *and bandwidth* h_N: *as* $N \to \infty$, $nN^{-1} \to \pi \in (0, 1)$, $h_N \to 0$ *and* $Nh_N^2 / (\log\log N) \to \infty$.
- **A6** *Inclusion probabilities* π_k *and* π_{kl}: *for all* N, $\min_{k \in U_N} \pi_k \geq \lambda > 0$, $\min_{k,\, l \in U_N} \pi_{kl} \geq \lambda^* > 0$ *and*
$$\limsup_{N \to \infty} n \max_{k,\, l \in U_{N:\, i \neq j}} |\pi_{kl} - \pi_k \pi_l| < \infty.$$
- **A7** *Additional assumptions involving higher-order inclusion probabilities*:
$$\lim_{N \to \infty} n^2$$
$$\max_{(k_1, k_2, k_3, k_4) \in D_{4,\, N}} |E_p(I_{k_1} - \pi_{k_1})(I_{k_2} - \pi_{k_2})(I_{k_3} - \pi_{k_3})(I_{k_4} - \pi_{k_4})|$$
$$< \infty,$$
where $D_{t,\, N}$ *denotes the set of all distinct* t*-tuples* $(k_1, k_2, \ldots, k_t)$ *from* U_N,
$$\lim_{N \to \infty}$$
$$\max_{(k_1,\, k_2,\, k_3,\, k_4) \in D_{4,\, N}} |E_p(I_{k_1} I_{k_2} - \pi_{k_1 k_2})(I_{k_3} I_{k_4} - \pi_{k_3 k_4})| = 0,$$
and
$$\limsup_{N \to \infty} n$$
$$\max_{(k_1,\, k_2,\, k_3) \in D_{3,\, N}} |E_p(I_{k_1} - \pi_{k_1})^2 (I_{k_2} - \pi_{k_2})(I_{k_3} - \pi_{k_3})| < \infty.$$
- **A8** *The matrix* $N^{-1} \boldsymbol{Z}_U^T (\boldsymbol{I} - \boldsymbol{S}_U^*) \boldsymbol{Z}_U$ *is invertible for all* N *with model probability* 1.

Assumption A8 is required so that the population estimator $\boldsymbol{B}$ is well-defined. The invertibility of the matrix in A8 depends on the combined effect of the bandwidth h and the joint distribution of the x_k and z_k. While it would in principle be possible to write down sufficient conditions for this, we opted for this simpler and more explicit approach.

Before giving the proofs of Theorems 3.1 and 3.2, we state and prove a number of lemmas.

Lemma 1 *Under the assumptions $A1$-$A7$,*

(a) *for all $k \in U$ and $d = 1, \ldots, D$,*

$$\frac{1}{N}\sum_U E_p(\boldsymbol{s}_{Ak}^T \boldsymbol{Y}_A - \boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U)^2 = O\left(\frac{1}{nh}\right)$$

and

$$\frac{1}{N}\sum_U E_p(\boldsymbol{s}_{Ak}^T \boldsymbol{Z}_{dA} - \boldsymbol{s}_{Uk}^T \boldsymbol{Z}_{dU})^2 = O\left(\frac{1}{nh}\right);$$

(b) *the $\boldsymbol{s}_{UK}^T \boldsymbol{Y}_U$ and $\boldsymbol{s}_{UK}^T \boldsymbol{Z}_U$ are uniformly bounded over all $k \in U$.*

Proof of Lemma 1: Since both the y_k and z_{dk} are bounded by assumption, part (a) can be shown using an identical reasoning as in Lemma 4 of Breidt and Opsomer (2000). While that lemma did not include a rate of convergence, that rate is readily derived by noting that

$$\frac{1}{N}\sum_{i,\,k \in U_N} z_{ik}^2 = O\left(\frac{1}{nh}\right)$$

in the notation of Breidt and Opsomer (2000) and then proceeding as in that proof.

Part (b) was proven directly in Lemma 2 (iv) of Breidt and Opsomer (2000).

Lemma 2 *Under assumptions $A1$-$A8$,*

$$\hat{\boldsymbol{B}} = \boldsymbol{B} + O_p(1/\sqrt{nh}),$$

with the rate holding component-wise, and $\boldsymbol{B}$ is bounded for all N.

Proof of Lemma 2: Write $\tilde{y}_k^{[s_U]} = \boldsymbol{s}_{Uk}^T \boldsymbol{Y}_U$ and $\tilde{y}_k^{[s_A]} = \boldsymbol{s}_{Ak} \boldsymbol{Y}_A$ for the population and sample smoothed versions of y_k, and similarly, $\tilde{\boldsymbol{z}}_k^{[s_U]} = \boldsymbol{s}_{Uk}^T \boldsymbol{Z}_U$ and $\tilde{\boldsymbol{z}}_k^{[s_A]} = \boldsymbol{s}_{Ak} \boldsymbol{Z}_A$. We rewrite expression (6) as a function of sample-weighted terms $\hat{\boldsymbol{t}}_l$, $l = 1, \ldots, 6$:

$$\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{t}_1 & \hat{\boldsymbol{t}}_2 \\ \hat{\boldsymbol{t}}_3 & \hat{\boldsymbol{t}}_4 \end{bmatrix}^{-1} \begin{bmatrix} \hat{t}_5 \\ \hat{\boldsymbol{t}}_6 \end{bmatrix},$$

where

$$\hat{t}_1 = \left(\frac{\hat{N}}{N}\right)^2$$

$$\hat{\boldsymbol{t}}_2 = \bar{\boldsymbol{z}}_\pi - \frac{1}{N}\sum_A \frac{\tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k}\left(1 - \frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_3 = \bar{\boldsymbol{z}}_\pi^T \left(\frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_4 = \frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T \boldsymbol{z}_k}{\pi_k} - \frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T \tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k} + \bar{\boldsymbol{z}}_\pi^T \frac{1}{N}\sum_A \frac{\tilde{\boldsymbol{z}}_k^{[s_A]}}{\pi_k}$$

$$\hat{t}_5 = \bar{y}_\pi - \frac{1}{N}\sum_A \frac{\tilde{y}_k^{[s_A]}}{\pi_k}\left(1 - \frac{\hat{N}}{N}\right)$$

$$\hat{\boldsymbol{t}}_6 = \frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T y_k}{\pi_k} - \frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T \tilde{y}_k^{[s_A]}}{\pi_k} + \bar{\boldsymbol{z}}_\pi^T \frac{1}{N}\sum_A \frac{\tilde{y}_k^{[s_A]}}{\pi_k}.$$

The sample-weighted estimator $\hat{\boldsymbol{B}}$ will be expanded around

$$\boldsymbol{B} = \begin{bmatrix} 1 & \bar{\boldsymbol{z}}_N \\ \bar{\boldsymbol{z}}_N^T & \boldsymbol{t}_4 \end{bmatrix}^{-1} \begin{bmatrix} \bar{y}_N \\ \boldsymbol{t}_6 \end{bmatrix}, \tag{15}$$

where

$$\boldsymbol{t}_4 = \frac{1}{N}\sum_U \boldsymbol{z}_k^T \boldsymbol{z}_k - \frac{1}{N}\sum_U \boldsymbol{z}_k^T \tilde{\boldsymbol{z}}_k^{[s_U]} + \bar{\boldsymbol{z}}_N^T \frac{1}{N}\sum_U \tilde{\boldsymbol{z}}_k^{[s_U]}$$

$$\boldsymbol{t}_6 = \frac{1}{N}\sum_U \boldsymbol{z}_k^T y_k - \frac{1}{N}\sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]} + \bar{\boldsymbol{z}}_N^T \frac{1}{N}\sum_U \tilde{y}_k^{[s_U]}$$

and the remaining $\boldsymbol{t}_l$ can be found in (15). The existence and continuity of the derivatives of $\hat{\boldsymbol{B}}$ with respect to the $\hat{\boldsymbol{t}}_l$ and evaluated at $\boldsymbol{t}_l$ follow from Lemma 1(b) and the existence of the inverse in (15), which is assumed by A8.

The result will follow from a 0th order Taylor expansion if we can show that $\hat{\boldsymbol{t}}_l - \boldsymbol{t}_l = O_p(1/\sqrt{nh})$ for all l (*e.g.*, Fuller (1996), Corollary 5.1.5). For $\hat{t}_1$ and $\hat{\boldsymbol{t}}_3$, this follows directly from A2 and A6. The remaining terms contain sums involving smoothed quantities $\tilde{\boldsymbol{z}}_k^{[s_A]}$ and $\tilde{y}_k^{[s_A]}$. We demonstrate the reasoning for one of those terms in $\hat{\boldsymbol{t}}_6$. We have

$$\frac{1}{N}\sum_A \frac{\boldsymbol{z}_k^T \tilde{y}_k^{[s_A]}}{\pi_k} - \frac{1}{N}\sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]} = \frac{1}{N}\sum_U \boldsymbol{z}_k^T \tilde{y}_k^{[s_U]}\left(\frac{I_k}{\pi_k} - 1\right)$$
$$+ \frac{1}{N}\sum_U \boldsymbol{z}_k^T (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]})\frac{I_k}{\pi_k},$$

and the first term is $O_p(1/\sqrt{n})$ by A6 and Lemma 1(b), using the same argument as in Lemma 4 of Breidt and Opsomer (2000). For the second term, use Schwarz's inequality

$$\left|\frac{1}{N}\sum_U \boldsymbol{z}_k^T (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]})\frac{I_k}{\pi_k}\right|$$
$$\leq \sqrt{\frac{1}{N}\sum_U \boldsymbol{z}_k^{[2]T} \frac{I_k}{\pi_k^2}}\sqrt{\frac{1}{N}\sum_U (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]})^2},$$

where $\boldsymbol{z}_k^{[2]}$ denotes that the squares are computed component-wise. The first term is bounded by A2 and A6, and the second term is $O_p(1/\sqrt{nh})$ by Lemma 1(a) and Markov's inequality. The desired result then follows by applying the same reasoning to the remaining terms in $\hat{\boldsymbol{t}}_2,\ \hat{\boldsymbol{t}}_4,\ \hat{t}_5,\ \hat{\boldsymbol{t}}_6$.

The boundedness of $\boldsymbol{B}$ follows directly from assumption A8, Lemma 1(b) and the boundedness of the z_k.

Lemma 3 *Under the assumptions A*1-*A*8, *we have*

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof of Lemma 3: Given expression (9), we need to show that

$$(\bar{\boldsymbol{z}}_N - \bar{\boldsymbol{z}}_\pi)(\boldsymbol{B} - \hat{\boldsymbol{B}}) = o_p\left(\frac{1}{\sqrt{n}}\right) \qquad (16)$$

$$\frac{1}{N}\sum_U (m_k - \hat{m}_k)\left(1 - \frac{I_k}{\pi_k}\right) = o_p\left(\frac{1}{\sqrt{n}}\right). \qquad (17)$$

Lemma 2 and assumptions A2, A5 and A6 show that $(\bar{\boldsymbol{z}}_N - \bar{\boldsymbol{z}}_\pi)(\boldsymbol{B} - \hat{\boldsymbol{B}}) = O_p(1/nh)$. In order to prove (17), we can rewrite it as

$$\begin{aligned}\frac{1}{N}\sum_U (m_k - \hat{m}_k)\left(1 - \frac{I_k}{\pi_k}\right) = {} & \frac{1}{N}\sum_U (\tilde{y}_k^{[s_U]} - \tilde{y}_k^{[s_A]})\left(1 - \frac{I_k}{\pi_k}\right) \\ & - \frac{1}{N}\sum_U (\tilde{\boldsymbol{z}}_k^{[s_U]} - \tilde{\boldsymbol{z}}_k^{[s_A]})\left(1 - \frac{I_k}{\pi_k}\right)\boldsymbol{B} \\ & - \frac{1}{N}\sum_U \tilde{\boldsymbol{z}}_k^{[s_A]}\left(1 - \frac{I_k}{\pi_k}\right)(\boldsymbol{B} - \hat{\boldsymbol{B}}).\end{aligned}$$

The first term on the right hand side has been proven to be $o_p(1/\sqrt{n})$ in Lemma 5 of Breidt and Opsomer (2000); this same Lemma and boundness of $\boldsymbol{B}$ provide the same rate for the second term. Assumptions A5-A6, Lemma 1(b) and Lemma 2 show that the third term is $O_p(1/n\sqrt{h})$ and the desired rate is achieved.

Lemma 4 *Under assumptions A*6 *and A*8,

$$\begin{aligned} E_p(\hat{y}_{\text{dif}}) &= \bar{y}_N \\ \text{Var}_p(\hat{y}_{\text{dif}}) &= \frac{1}{N^2}\sum\sum_{k,l\in U} (\pi_{kl} - \pi_k\pi_l)\frac{y_k - g_k}{\pi_k}\frac{y_l - g_l}{\pi_l} \\ &= O\left(\frac{1}{n}\right).\end{aligned}$$

Proof of Lemma 4: The properties of the difference estimator are readily computed. The rate of the design variance follows from the stated assumptions using the same reasoning as in Lemma 4 of Breidt and Opsomer (2000).

Lemma 5 *Under assumptions A*1-*A*8,

$$\hat{V}(\hat{y}_{\text{reg}}) = \text{Var}_p(\hat{y}_{\text{dif}}) + o_p\left(\frac{1}{n}\right).$$

Proof of Lemma 5: The reasoning for this proof will closely follow that of Theorem 3 of Breidt and Opsomer (2000). We write

$$\begin{aligned}\hat{V}(\hat{y}_{\text{reg}}) - \text{Var}_p(\hat{y}_{\text{dif}}) = {} & (\hat{V}(\hat{y}_{\text{reg}}) - \hat{V}(\hat{y}_{\text{dif}})) \\ & + (\hat{V}(\hat{y}_{\text{dif}}) - \text{Var}_p(\hat{y}_{\text{dif}})) \quad (18)\end{aligned}$$

with

$$\hat{V}(\hat{y}_{\text{dif}}) = \frac{1}{N^2}\sum\sum_A \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k - g_k}{\pi_k}\frac{y_l - g_l}{\pi_l}.$$

Since

$$\frac{1}{N}\sum_U (y_k - g_k)^4 < \infty.$$

by assumptions A1-A3 and from Lemmas 1(b) and 2, the approach used for the term A_N of Breidt and Opsomer (2000) can be used to show that

$$\text{E}_p\,|\hat{V}(\hat{y}_{\text{dif}}) - \text{Var}_p(\hat{y}_{\text{dif}})| = o\left(\frac{1}{n}\right),$$

which provides the desired consistency by the Markov inequality.

For the first term in (18), note that

$$\begin{aligned}\hat{g}_k - g_k = {} & (\tilde{y}_k^{[s_A]} - \tilde{y}_k^{[s_U]}) - (\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]})\ (\hat{\boldsymbol{B}} - \boldsymbol{B}) \\ & + (\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_k^{[s_U]})(\hat{\boldsymbol{B}} - \boldsymbol{B}) - (\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]})\boldsymbol{B},\end{aligned}$$

so that

$$(\hat{V}(\hat{y}_{\text{reg}}) - \hat{V}(\hat{y}_{\text{dif}})) =$$

$$\frac{1}{N^2}\sum\sum_U \left\{\begin{aligned} & -2\frac{y_k - g_k}{\pi_k}\frac{\hat{g}_l - g_l}{\pi_l} \\ & \quad + \frac{\hat{g}_k - g_k}{\pi_k}\frac{\hat{g}_l - g_l}{\pi_l}\end{aligned}\right\}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} I_k I_l$$

can be decomposed into variance terms involving sample and population smooths and parameter estimators. Each of these terms can be shown to be $o_p(1/n)$. We demonstrate the approach on one of the terms:

$$\begin{aligned} & \left|\frac{1}{N^2}\sum\sum_U \frac{y_k - g_k}{\pi_k}\frac{\hat{\boldsymbol{z}}_l - \boldsymbol{z}_l}{\pi_l}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} I_k I_l(\hat{\boldsymbol{B}} - \boldsymbol{B})\right| \\ & \le \left(\frac{C_1}{N} + C_2 \max|\pi_{kl} - \pi_k\pi_l|\right)\frac{1}{N}\sum_U |\tilde{\boldsymbol{z}}_k^{[s_A]} - \tilde{\boldsymbol{z}}_k^{[s_U]}||\hat{\boldsymbol{B}} - \boldsymbol{B}| \\ & = o_p\left(\frac{1}{n}\right)\end{aligned}$$

where $C_1, C_2 < \infty$ summarize the bounded terms (by assumptions A1-A3 and A6 and Lemma 1(b)), and the rate of convergence is the result of assumption A6 and Lemmas 1(a) and 2.

Proof of Theorem 3.1: In Lemma 3, we show that

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where $\hat{y}_{\text{dif}}$ is the difference estimator (3). The result immediately follows from assumption A5 and Lemma 4.

Proof of Theorem 3.2: Note that $\hat{y}_{\text{dif}}$ can be written as the sum of a population constant and an expansion estimator of the form $\bar{y}_\pi$ by defining a new variable $y_k - \mathbf{s}_{Uk}^T \mathbf{Y}_U + \mathbf{s}_{Uk}^T \mathbf{Z}_U \mathbf{B} - \mathbf{z}_k \mathbf{B}$ for $k \in U$. As is the case for the original y_k, this new variable has bounded support by Lemma 1(b) and a variance of order $O(1/n)$ by Lemma 4. Hence, existence of the CLT for $\bar{y}_\pi$ implies existence of the CLT for $\hat{y}_{\text{dif}}$. Also, $\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p(1/\sqrt{n})$ by Lemma 3, so that $\sqrt{n}\,\hat{y}_{\text{reg}}$ and $\sqrt{n}\,\hat{y}_{\text{dif}}$ have the same asymptotic distribution. Applying Slutsky's Theorem and Lemma 5 complete the proof.

References

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

Fuller, W.A. (1996). *Introduction to Statistical Time Series* (2nd Ed.). New York: John Wiley & Sons, Inc.

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models*. Washington, D.C.: Chapman and Hall.

Isaki, C., and Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Larsen, D.P., Thornton, K.W., Urquhart, N.S. and Paulsen, S.G. (1993). Overview of survey design and lake selection. EMAP - Surface Waters 1991 Pilot Report. Technical Report EPA/620/R - 93/003, U.S. Environmental Protection Agency. (Eds. D.P Larsen and S.J. Christie).

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73, 166-179.

Opsomer, J.D., Breidt, F.J., Moisen, G.G. and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*. To appear.

Opsomer, J.-D., and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25, 186-211.

Opsomer, J.D., and Ruppert, D. (1999). A root-n consistent estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics*, 8, 715-732.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Speckman, P.E. (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Society*, Series B, 50, 413-436.

Stoddard, J.L., Kahl, J.S., Deviney, F.A., DeWalle, D.R., Driscoll, C.T., Herlihy, A.T., Kellogg, J.H., Murdoch, P.S., Webb, J.R. and Webster, K.E. (2003). Response of surface water chemistry to the Clean Air Act Amendments of 1990. Technical Report EPA/620/R-03/001, U.S. Environmental Protection Agency, Washington, DC.

U.S. National Acid Precipitation Assessment Program (1991, November). 1990 Integrated Assessment Report. Technical report, Washington, DC.