Model-Assisted Estimation for Complex Surveys Using Penalized Splines

F.J. Breidt

Colorado State University*

G. Claeskens

Katholieke Universiteit Leuven†

J.D. Opsomer

Iowa State University‡

March 23, 2005

Abstract

Estimation of finite population totals in the presence of auxiliary information is considered. A class of estimators based on penalized spline regression is proposed. These estimators are weighted linear combinations of sample observations, with weights calibrated to known control totals. Further, they allow straightforward extensions to multiple auxiliary variables and to complex designs. Under standard design conditions, the estimators are design consistent and asymptotically normal, and they admit consistent variance estimation using familiar design-based methods. Data-driven penalty selection

*Colorado State University, Fort Collins, CO 80523 USA; jbreidt@stat.colostate.edu.

†K. U. Leuven, Naamsestraat 69, 3000 Leuven, Belgium; Gerda.Claeskens@econ.kuleuven.ac.be.

‡Iowa State University, Ames, IA, 50011 USA; jopsomer@iastate.edu.

is considered in the context of unequal probability sampling designs. Simulation experiments show that the estimators are more efficient than parametric regression estimators when the parametric model is incorrectly specified, while being approximately as efficient when the parametric specification is correct. An example using Forest Health Monitoring survey data from the U.S. Forest Service demonstrates the applicability of the methodology in the context of a two-phase survey with multiple auxiliary variables.

Key Words: calibration, generalized regression estimation, nonparametric regression, restricted maximum likelihood.

1 Introduction

Large-scale surveys are routinely conducted by both governmental and private organizations. We consider estimation for lichen data collected in the *Forest Health Monitoring* (FHM) survey conducted in Utah by the U.S. Forest Service (Rogers et al. 2001). In addition to the FHM variables, large amounts of auxiliary information are available for the region of interest. Improving the efficiency of survey estimators in situations like the FHM can be achieved by taking advantage of this auxiliary information.

An important area of methodological research is to improve the precision of estimators, while continuing to rely on the sampling design as the primary probability generating mechanism. We propose a new estimator that follows this "design-based" paradigm, but uses a class of modelling techniques that is more flexible than those currently in use.

More formally, we consider the classical finite population estimation problem, in which a population U of size N is surveyed according to sampling design $p(\cdot)$. Let y_i represent a generic characteristic for the ith element of the population. We focus on the estimation of $t_y = \sum_U y_i$, the total of the y_i for the population, based on observing the y_i for a

sample of population elements $s \subset U$, selected with probability $p(s)$. Let $\pi_i = \Pr(i \in s) = \sum_{s:i \in s} p(s)$ denote the *inclusion probability* for element i under the design $p(\cdot)$. If no information other than the inclusion probabilities is used to estimate t_y, the Horvitz-Thompson estimator

$$\widehat{t}_{y,\mathrm{HT}} = \sum_s \frac{y_i}{\pi_i} \tag{1}$$

(Horvitz and Thompson, 1952) is frequently used. The closely related Hájek estimator (Hájek, 1971), is a common alternative in unequal probability sampling situations.

Let x_i represent a characteristic for the ith element that is known for all $i \in U$ (for now, we assume that x is a univariate and continuous variable and that we have a single phase of sampling, but we will relax those restrictions in later sections). If the x_i and y_i are related to each other, an estimator that takes advantage of that fact is likely to be more efficient than (1). A double goal of high efficiency and design consistency is the motivation for estimators such as the classical ratio and regression estimators (e.g. Cochran, 1977), as well as for the generalized regression estimator (GREG) described in Cassel et al. (1977) or the calibration estimator (Deville and Särndal, 1992), among others. Särndal et al. (1992) provide a comprehensive description of the *model-assisted* framework for constructing survey estimators.

In Breidt and Opsomer (2000), a nonparametric model-assisted regression estimator was proposed, which allows the relationship between the variables to be any smooth function, not just one that belongs to a pre-specified parametric family. Breidt and Opsomer (2000) used kernel-based *local polynomial regression* (see Wand and Jones, 1995), and showed that the nonparametric estimator shares the asymptotic design properties of the parametric model-assisted estimators. Nonparametric regression can significantly improve the efficiency of the estimators when parametric models are misspecified. Breidt and Opsomer (2003) extend this approach from the single-covariate model to the case of

the semiparametric additive model.

In the current paper, we investigate the properties of nonparametric model-assisted estimators when penalized regression splines are used. Eilers and Marx (1996) introduced penalized spline smoothers (often called *P-splines*), and further developments and results on these low-rank smoothers have been obtained by Ruppert and Carroll (2000) and Wand (2003). For the current state of the art on this estimation method, we refer to Ruppert et al. (2003). Recently, Zheng and Little (2003) have proposed a model-based survey estimator that uses penalized splines to account for the effect of non-ignorable design weights, and have extended it to the case of two-stage sampling in Zheng and Little (2004).

P-splines offer a number of advantages over kernel-based methods that make them an attractive smoothing method in the model-assisted context. Incorporating multiple covariates as well as combinations of categorical variables, parametric and non-parametric terms, is straightforward, as shown in Aerts et al. (2002). Another important advantage is the relative ease with which P-spline estimators can be computed, even for large datasets or datasets with regions of sparse data. Finally, an important practical consideration is that, since they are more closely related to parametric models, estimators based on spline models are easier to implement in existing survey estimation procedures. This will be further discussed in later sections.

2 Penalized spline model-assisted estimator

2.1 Definition of the estimator

To introduce the estimator, consider the superpopulation regression model

$$y_i = m(x_i) + \varepsilon_i, \tag{2}$$

where the ε_i are independent random variables with mean zero and variance $v(x_i)$. In order to develop the model-assisted estimator, we treat $\{(x_i, y_i) : i \in U\}$ as a realization from this superpopulation model. If this entire realization were observed, we could define a P-spline estimator for the function $m(\cdot)$ as follows. Let

$$m(x;\beta) = \beta_0 + \beta_1 x + \cdots + \beta_q x^q + \sum_{k=1}^{K} \beta_{q+k}(x - \kappa_k)_+^q, \tag{3}$$

where $(t)_+^q = t^q$ if $t > 0$ and 0 otherwise, q is the degree of the spline, and the κ_k are called the *knots*. Typically, q is kept fixed and low, e.g. $q \leq 3$; $q = 3$ corresponds to cubic splines, a popular choice in practice. Flexibility of the model is obtained by means of the truncated qth degree polynomials, denoted by the last sum in (3). The locations of the terms in that sum are determined by the choice of the fixed knots $\kappa_1 < \cdots < \kappa_K$. Both the number K and the location of the knots are typically assumed fixed in the P-spline regression literature, and we will do the same. The knots κ_k $(k = 1, \ldots, K)$ can, for example, be uniformly spread sample quantiles of the unique x_i values. That is, κ_k equals the $k/(K+1)$th sample quantile. This default choice seems to perform well in practice.

The vector $\beta = (\beta_0, \ldots, \beta_{q+K})^T$ is the coefficient vector. If the number of knots K is sufficiently large, the class of functions $m(x;\beta)$ is very large and can approximate most smooth functions with a high degree of accuracy. Note also that, in the model-assisted approach to survey inference, the superpopulation model does not have to fit the data exactly for it to be a useful way to construct a survey estimator.

For ease of exposition we wrote (3) using a truncated polynomial basis, as is also done in Ruppert et al. (2003). For any given degree q, the resulting spline fit can equivalently be represented by using other basis functions such as a q-degree B-spline basis (Eilers

and Marx, 1996) or the Demmler-Reinsch basis (Nychka and Cummins, 1996), which are more efficient from a computational standpoint. The truncated polynomial basis makes it easier to interpret the results, so we will focus on that basis.

Too many knots results in an unstable and highly variable fit, while too few knots can lead to bias. While knot selection methods exist, P-spline regression relies instead on a different approach to ensure a reasonable fit: the number of knots K is chosen to be large, but the influence of the knots is limited by putting a constraint on the size of the spline coefficients. A typical constraint in the truncated polynomial case is to bound $\sum_{k=1}^{K} \beta_{q+k}^2$ by some constant, while leaving the polynomial coefficients $\beta_0, \ldots, \beta_q$ unconstrained. When this constraint is incorporated into the least squares criterion via a Lagrange multiplier, the population estimator for β is defined as the minimizer of

$$\sum_{i \in U} (y_i - m(x_i; \beta))^2 + \alpha \sum_{k=1}^{K} \beta_{q+k}^2 \tag{4}$$

for some fixed constant $\alpha \geq 0$. The smoothness of the resulting fit depends on the value chosen for α, with larger values corresponding to smoother fits. Choice of α will be further discussed below.

Let X represent the matrix with rows $X_i^T = \{1, x_i, \ldots, x_i^q, (x_i - \kappa_1)_+^q, \ldots, (x_i - \kappa_K)_+^q\}$ for $i \in U$, and let Y denote the column vector of response values y_i for $i \in U$. Define the diagonal matrix $A_\alpha = \text{diag}\{0, \ldots, 0, \alpha, \ldots, \alpha\}$, with $q+1$ zeros on the diagonal followed by K penalty constants α, corresponding to the K truncated polynomial terms in (3). If the population U is fully observed, the penalized least squares estimator for the coefficient vector of (3) has the ridge-regression representation:

$$B_U = (X^T X + A_\alpha)^{-1} X^T Y. \tag{5}$$

Other penalty structures besides the one in (4) are possible, and in those cases, A_α might no longer be diagonal. For instance when B-splines are used instead of polynomial splines,

the columns of X contain the B-spline basis functions only (i.e. there is no unconstrained polynomial part as in (3)), and A_α is typically a banded matrix. However, expression (5) will continue to hold for any penalized spline formulation. Let $m_i = m(x_i; B_U) \equiv X_i^T B_U, i \in U$ denote the P-spline fit obtained from this hypothetical population fit at x_i. If these fitted values are known, they can be incorporated into the survey estimation by constructing the *difference estimator*

$$\widehat{t}_{y,\text{diff}} = \sum_U m_i + \sum_s \frac{y_i - m_i}{\pi_i} \tag{6}$$

(Särndal et al. 1992, p.221). Other approaches are also possible. For example, the design properties of a prediction-based P-spline estimator as in Zheng and Little (2003) will be evaluated via simulation in §3.1.

The difference estimator is design unbiased and its design variance is

$$\text{Var}_p(\widehat{t}_{y,\text{diff}}) = \sum_U \sum_U (\pi_{ij} - \pi_i \pi_j) \frac{y_i - m_i}{\pi_i} \frac{y_j - m_j}{\pi_j} \tag{7}$$

where $\pi_{ij} = \Pr(i \in s, j \in s)$, the joint inclusion probability for elements $i, j \in U$. As is clear from expression (7), the efficiency of $\widehat{t}_{y,\text{diff}}$ depends on how well the model function $m(x_i; \beta)$ (or, more accurately, its population fit m_i) approximates the variable y_i. If the assumed superpopulation model is an appropriate representation for the population, then $\widehat{t}_{y,\text{diff}}$ will be more efficient than $\widehat{t}_{y,\text{HT}}$.

The estimator (6) is infeasible, since the m_i cannot be calculated. However, given a sample s, the m_i in (6) can be replaced by sample-based estimators, denoted by $\widehat{m}_i$ and constructed as follows. Define the diagonal matrix of inverse inclusion probabilities $W = \text{diag}_{j \in U}\{1/\pi_j\}$ and its sample submatrix $W_s = \text{diag}_{j \in s}\{1/\pi_j\}$. Similarly, let X_s be the submatrix of X consisting of those rows for which $j \in s$. For fixed α and under suitable regularity conditions, the π-weighted estimator

$$\widehat{\beta} = (X_s^T W_s X_s + A_\alpha)^{-1} X_s^T W_s Y_s = G_\alpha Y_s \tag{8}$$

is a design-consistent estimator of B_U in (5). Notice that the penalty matrix is identical to the one in the population fit (5), and that an unweighted version can be obtained by setting W equal to the identity matrix I.

Define $\widehat{m}_i = m(x_i, \widehat{\beta}) \equiv X_i^T \widehat{\beta}$. The model-assisted P-spline estimator is then defined as

$$\widehat{t}_{y,\text{spl}} = \sum_U \widehat{m}_i + \sum_s \frac{y_i - \widehat{m}_i}{\pi_i}. \tag{9}$$

The design properties of $\widehat{t}_{y,\text{spl}}$ are explored in the next section.

2.2 Properties of the estimator

The close relationship between penalized spline regression and linear least squares regression has a number of important practical implications. Introducing the indicator function $I_i = 1$ if $i \in s$ and $I_i = 0$ otherwise, and the indicator vector e_i which is a zero vector except for an entry of one at position i, we can rewrite (9) as

$$\widehat{t}_{y,\text{spl}} = \sum_{i \in s} \left\{ \pi_i^{-1} + \sum_{j \in U} (1 - I_j/\pi_j) X_j^T G_\alpha e_i \right\} y_i \equiv \sum_s w_{i(s)} y_i \tag{10}$$

which shows that $\widehat{t}_{y,\text{spl}}$ is a linear estimator, making it useful in a survey estimation context because survey weights $w_{i(s)}$ are obtained. Chambers (1996) proposed a regression estimator of a form similar to (9) for the case of linear regression on multiple auxiliary variables, with a ridge matrix $\text{diag}(\alpha_1, \ldots, \alpha_p)$, where $\alpha_i = 0$ for those covariates (e.g., the intercept) corresponding to calibration constraints that must be met. This formulation avoids negative regression weights. In the P-spline regression case, the penalty α could also be adjusted if negative weights occur, but our main purpose for using a "ridged" parameter estimator as in (8) is to control the smoothness of the spline fit.

The P-spline estimator shares many of the good practical properties of generalized regression (GREG) estimators in common use in large-scale survey estimation. It is

readily checked that, as long as an intercept is included in the model, the estimator (9) shares the properties of location and scale invariance of the GREG, in the sense that $\sum_s w_{i(s)}(ay_i + b) = a\ \widehat{t}_{y,spl} + Nb$ for any fixed a, b. Like the ridge regression estimator in equation (10) of Chambers (1996), the P-spline estimator (9) is calibrated for the parametric portion of the model; that is, $\sum_s w_{i(s)} x_i^d = \sum_U x_i^d$ for $0 \leq d \leq q$. This calibration can also be made to hold for any of the truncated polynomial terms in (3), as long as these terms are left unpenalized. These invariance and calibration properties were also shared by the model-assisted local polynomial estimator of Breidt and Opsomer (2000).

There are a number of additional properties shared by the P-spline and GREG estimators, which do not hold for the local polynomial estimator. Let $t_X = \sum_U X_i$ and let $\widehat{t}_{X,HT} = \sum_s X_i \pi_i^{-1}$ denote its Horvitz-Thompson estimator. Then, $\widehat{t}_{y,\text{spl}} = \widehat{t}_{y,HT} + (t_X - \widehat{t}_{X,HT})^T \widehat{\beta}$, the same form as the GREG estimator. Further, with $e_1 = (1, 0, \ldots, 0)^T$, we have $e_1^T X_i = 1$, $e_1^T A_\alpha = (0, 0, \ldots, 0)$, and

$$\widehat{t}_{X,HT}^T \widehat{\beta} = \left(\sum_{i \in s} \frac{e_1^T X_i X_i^T}{\pi_i} + e_1^T A_\alpha \right) \left(\sum_{j \in s} \frac{X_j X_j^T}{\pi_j} + A_\alpha \right)^{-1} \sum_{j \in s} \frac{X_j y_j}{\pi_j} = \widehat{t}_{y,HT}$$

so the P-spline estimator can be written in the simple form $\widehat{t}_{y,\text{spl}} = t_X^T \widehat{\beta} = \sum_U \widehat{m}_i$, a property it shares with both ratio and linear regression estimators (see Särndal et al. 1992, p. 231). Note from this representation that $\widehat{t}_{y,\text{spl}}$ can be computed using only sample x_i's along with population counts and totals in strata defined by the knots, $\sum_{i \in U} I_{\{x_i > \kappa_k\}}$ and $\sum_{i \in U} x_i I_{\{x_i > \kappa_k\}}$. By contrast, classical ratio or regression needs only sample x_i's and population size and x-total, while the local polynomial estimator in Breidt and Opsomer (2000) needs all population x_i's. Finally, though both P-spline and kernel-based model-assisted survey regression estimators have GREG representations, P-splines make it much easier to incorporate nonparametric regression into survey estimation procedures. In

particular, they readily allow for parametric covariates, including categorical terms, and for additional nonparametric covariates.

For the remaining design properties, we will use the traditional finite population asymptotic framework, in which both the population U and the sampling design $p(\cdot)$ are embedded into a sequence of such populations and designs indexed by N, $\{U_N, p_N(\cdot)\}$, with $N \to \infty$. The $o_p(\cdot)$ and $O_p(\cdot)$ notation below is with respect to this sequence of designs. See Isaki and Fuller (1982) for an early version of this asymptotic framework. For simplicity, we will restrict our attention to the case in which the sample size, denoted by n_N, is fixed for each N, and we also assume that $n_N \to \infty$. As above, q, K and the $\{\kappa_k\}$ are fixed. In order to prove our results, we make the following technical assumptions.

A1. $B = \lim_{N\to\infty} B_U$ *exists, and* $\widehat{\beta} - B_U = o_p(1)$.

A2. *The limiting design covariance matrix of the normalized Horvitz-Thompson estimators,*

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{yx}^T & \Sigma_{xx} \end{bmatrix} = \lim_{N\to\infty} n_N \begin{bmatrix} \sum\sum_U \frac{\pi_{ij}-\pi_i\pi_j}{N^2}\frac{y_i y_j}{\pi_i\pi_j} & \sum\sum_U \frac{\pi_{ij}-\pi_i\pi_j}{N^2}\frac{y_i X_j^T}{\pi_i\pi_j} \\ \sum\sum_U \frac{\pi_{ij}-\pi_i\pi_j}{N^2}\frac{X_i y_j}{\pi_i\pi_j} & \sum\sum_U \frac{\pi_{ij}-\pi_i\pi_j}{N^2}\frac{X_i X_j^T}{\pi_i\pi_j} \end{bmatrix},$$

is positive definite.

A3. *The normalized Horvitz-Thompson estimators satisfy a central limit theorem:*

$$\frac{\sqrt{n_N}}{N}\begin{bmatrix} \sum_U y_i\left(\frac{I_i}{\pi_i}-1\right) \\ \sum_U X_i^T\left(\frac{I_i}{\pi_i}-1\right) \end{bmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0,\Sigma).$$

A4. *The estimated covariance matrix for the Horvitz-Thompson estimators is design consistent in the following sense:*

$$n_N \begin{bmatrix} \sum\sum_s \frac{\pi_{ij}-\pi_i\pi_j}{N^2\pi_{ij}}\frac{y_i y_j}{\pi_i\pi_j} & \sum\sum_s \frac{\pi_{ij}-\pi_i\pi_j}{N^2\pi_{ij}}\frac{y_i X_j^T}{\pi_i\pi_j} \\ \sum\sum_s \frac{\pi_{ij}-\pi_i\pi_j}{N^2\pi_{ij}}\frac{X_i y_j}{\pi_i\pi_j} & \sum\sum_s \frac{\pi_{ij}-\pi_i\pi_j}{N^2\pi_{ij}}\frac{X_i X_j^T}{\pi_i\pi_j} \end{bmatrix} - \Sigma = o_p(1)$$

as $N \to \infty$.

Remark 1. Assumption A1 ensures that the sample fit $\widehat{\beta}$ and finite population fit B_U share a common limit. A1 is weaker than assuming model (2) with (3) as its true mean function. The assumption depends on the distribution of the x_i, as well as on the knots and the penalty constant α. In Breidt and Opsomer (2000), existence of the population nonparametric fit was achieved by a combination of assumptions on the density of the x_i and the bandwidth h_N. Similarly, sufficient conditions for A1 could be constructed by further specifying the distribution of the x_i, the placement of the knots and the size of the penalty α. This will not be further pursued here.

Remark 2. Assumptions A2 and A3 are satisfied for commonly used fixed sample size designs in reasonably behaved finite populations. They will also hold for sampling designs with random sample size, but in that case n_N is replaced by $E_p(n_N)$ in the normalization. Assumption A4 is satisfied by many but not all common designs. For example, A4 would not hold for systematic sampling or one-per-stratum designs.

The following results establish design consistency and asymptotic normality of $\widehat{t}_{y,spl}$, along with a consistent variance estimator. Proofs are immediate from the given assumptions and are omitted.

Theorem 2.1. *Under assumptions A1–A3, the penalized spline estimator* $\widehat{t}_{y,spl}$ *is design* $\sqrt{n_N}$*-consistent for* t_y*, in the sense that* $N^{-1}\left(\widehat{t}_{y,\mathrm{spl}} - t_y\right) = O_p\left(n_N^{-1/2}\right)$*. Further,* $\left(Var_p(\widehat{t}_{y,\mathrm{diff}})\right)^{-1/2}\left(\widehat{t}_{y,\mathrm{spl}} - t_y\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ *with* $Var_p(\widehat{t}_{y,\mathrm{diff}})$ *given in* (7)*.*

Theorem 2.2. *Under assumptions A1, A2 and A4,*

$$\widehat{V}(\widehat{t}_{y,\mathrm{spl}}) = \sum_s \sum_s \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \frac{y_i - \widehat{m}_i}{\pi_i} \frac{y_j - \widehat{m}_j}{\pi_j} = \mathrm{Var}_p(\widehat{t}_{y,\mathrm{diff}}) + o_p\left(\frac{N^2}{n_N}\right).$$

Corollary 2.1. *Under assumptions A1–A4, the penalized spline estimator* $\widehat{t}_{y,spl}$ *satisfies* $\left(\widehat{V}(\widehat{t}_{y,\mathrm{spl}})\right)^{-1/2}\left(\widehat{t}_{y,\mathrm{spl}} - t_y\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

These theoretical results show that it is possible to use a flexible model specification as in (3) for the superpopulation model, while maintaining the same basic design properties as have been obtained for much more restrictive model specifications in the ratio or linear regression estimators. Unlike the theory for the kernel-based estimators in Breidt and Opsomer (2000), the theory derived above for the P-spline estimator closely follows that used for traditional survey (linear) regression estimation.

2.3 Spline regression with data-driven penalty

The previous results all assume that the various spline settings, including the number and placement of the knots, the type and degree of the spline basis functions and the penalty constant α, are all determined and fixed before the model is fitted. Naturally, the efficiency in the estimation of a given population quantity is directly affected by these factors. Ruppert et al. (2003, chapter 5) observe that for standard nonparametric regression, it is sufficient to focus on the choice of α, since (within reasonable bounds) the choice of the remaining settings has limited effect on the resulting fits once α is allowed to vary; our experience in the finite population setting is consistent with this observation.

It should be noted, however, that in the survey context, trying to estimate the "optimal" penalty is not as practically relevant as in the model-based regression context, because an estimator is typically not constructed for a single variable, but rather for a (sometimes very large) *set* of variables collected during a survey. This is the reason why survey estimators are expressed as a weighted sum, so that the same weights can be applied to all the variables. A set of weights that are optimized for a single variable in that

survey might not work as well for other variables. Instead, a "compromise" choice for the penalty that works reasonably well for many variables in a survey would be a better choice in practice.

Nevertheless, data-driven selection of α could still be of some use when the survey has a small number of key variables, or more generally in the exploratory stages of weight construction for a survey, where different spline formulations can be tested on some variables. We therefore briefly describe how to adapt to the survey context an existing method for selecting α. This method is popular in the spline regression literature.

For data-driven selection of the value of α in the regression context, Ruppert et al. (2003) recommend treating the estimator (5) as resulting from a linear mixed model specification, and using *Restricted Maximum Likelihood* (REML) (Patterson and Thompson, 1971; Searle et al. 1992) to estimate an appropriate value for α. To do this, the estimator (5) is viewed as the population-level REML estimator for the model

$$y_i = f_i^T \beta_f + r_i^T u + \varepsilon_i \tag{11}$$

where $f_i^T = (1, x_i, \ldots, x_i^q)$ and β_f represent the (fixed) terms in (3) and their corresponding parameters, and $r_i^T = ((x_i - \kappa_1)_+^q, \ldots, (x_i - \kappa_K)_+^q)$ contains the truncated polynomial terms. The random vector u has zero mean and covariance matrix $\sigma_u^2 I_K$, and is independent of $\{\varepsilon_i\}$, which are independent random variables with mean zero and variance σ_ε^2. Let F and R represent the matrices with rows equal to f_i^T and r_i^T, respectively, so that $X = [\, F \; R \,]$. The population REML estimator for (11) is given by

$$B_{U,reml} = (X^T X + A_{\widehat{\alpha}})^{-1} X^T Y, \tag{12}$$

identical to the estimator in (5) but with $A_{\widehat{\alpha}} = \text{diag}\{0, \ldots, 0, \widehat{\alpha}, \ldots, \widehat{\alpha}\}$, $\widehat{\alpha} = \widehat{\sigma}_\varepsilon^2 / \widehat{\sigma}_u^2$, and

$\widehat{\sigma}^2_\varepsilon, \widehat{\sigma}^2_u$ solving the REML estimating equations

$$\begin{aligned} Y^T A(A^T V A)^{-1} A^T R R^T A(A^T V A)^{-1} A^T Y &= \mathrm{tr}((A^T V A)^{-1} A^T R R^T A) \\ Y^T A(A^T V A)^{-1} A^T A(A^T V A)^{-1} A^T Y &= \mathrm{tr}((A^T V A)^{-1} A^T A) \end{aligned} \tag{13}$$

where $V = \mathrm{Var}(Y) = \sigma^2_\varepsilon I_N + \sigma^2_u R^T R$ and A is a matrix such that $A^T F = 0$ and $\mathrm{rank}(A) = N - (p+1)$. These estimators can be calculated using existing software routines, including `proc mixed` in SAS or `lme()` in S-Plus.

Assuming that the estimator $\widehat{\alpha}$ is well-defined for the population (e.g. Jiang, 1996), we can define a design-based $\sqrt{n}$-consistent estimator for that quantity. We do this by constructing design-consistent estimators for all the terms in (13). Letting $\Pi_{2s} = [1/\pi_{kl}]_{k,l \in s}$, the sample-weighted estimating equations are given by

$$\begin{aligned} Y_s^T \left\{ \Pi_{2s} * A_s (A^T V A)^{-1} A^T R R^T A (A^T V A)^{-1} A_s^T \right\} Y_s &= \mathrm{tr}((A^T V A)^{-1} A^T R R^T A) \\ Y_s^T \left\{ \Pi_{2s} * A_s (A^T V A)^{-1} A^T A (A^T V) A)^{-1} A_s^T \right\} Y_s &= \mathrm{tr}((A^T V A)^{-1} A^T A) \end{aligned} \tag{14}$$

where A_s denotes the submatrix of A containing the n_N columns corresponding to the sampled elements of the population and $\{\cdot * \cdot\}$ denotes an element-wise product.

Under suitable regularity conditions which we do not further explore here, the solutions $\widehat{\sigma}^2_{\varepsilon,s}, \widehat{\sigma}^2_{u,s}$ to these equations should be $\sqrt{n}$-consistent for $\widehat{\sigma}^2_\varepsilon, \widehat{\sigma}^2_u$, the solutions to (13), using an approach similar to that in Wu (1999). It would then follow that $\widehat{\alpha}_s = \widehat{\sigma}^2_{\varepsilon,s} / \widehat{\sigma}^2_{u,s}$ is $\sqrt{n}$-consistent for $\widehat{\alpha}$. Hence, the spline regression estimator that uses the mixed model formulation to determine the value for the penalty α instead of treating it as fixed will follow the properties described in Theorems 2.1–2.2 and Corollary 2.1.

2.4 Extensions

We consider the semiparametric additive model as a first extension. This is defined as $y_i = \sum_{d=1}^D m_d(x_{di}) + z_i^T \gamma + \varepsilon_i$, where the x_d are covariates that are modelled using smooth

univariate functions, and $z^T = (z_1, \ldots, z_q)$ is a vector of covariates that is incorporated into the model through parametric terms. The latter terms would include all categorical variables as well as those for which element-level information outside the sample is not available, but for which population totals are known from some source outside the survey. We define the population estimators for the β_d and γ as the minimizers of

$$\sum_{i \in U} \left(y_i - \sum_{d=1} m_d(x_{di}; \beta_d) - z_i^T \gamma \right)^2 + \sum_{d=1}^{D} \alpha_d \sum_{k=1}^{K_d} \beta_{d,p+k}^2$$

for some fixed constants $\alpha_d, d = 1, \ldots, D$. With little modification, the theory from §2.2 continues to hold. We refer to Ruppert et al. (2003) for an overview of additional extensions such as multivariate spline models, locally varying penalties and robust regression.

Second, we briefly describe the extension to two-phase sampling here, since it will be used in §3.2 below. In the notation of Särndal et al. (1992, ch.9), let s_a denote the first phase sample with inclusion probabilities $\pi_{ai} = \Pr(i \in s_a)$, and $s \subseteq s_a$ the phase two sample, with conditional inclusion probabilities $\pi_{i|s_a} = \Pr(i \in s | i \in s_a)$. Suppose that we have the vector of auxiliary variables X_{1i} available for all $i \in U$, and $X_i^T = (X_{1i}^T, X_{2i}^T)$ for all $i \in s_a$. Let $\widehat{g}_i, i \in s_a$ denote the predicted values for the model relating the y_i to the X_i, and let $\widehat{g}_{1i}, i \in U$ denote the predicted values for the model relating the y_i to the X_{1i}. The model-assisted estimator is defined as

$$\widehat{t}_{y,\text{spl}} = \sum_{U} \widehat{g}_{1i} + \sum_{s_a} \frac{\widehat{g}_i - \widehat{g}_{1i}}{\pi_{ai}} + \sum_{s} \frac{y_i - \widehat{g}_i}{\pi_{ai}\, \pi_{i|s_a}}. \tag{15}$$

The properties of this estimator can be derived using the approach of Särndal et al. (1992) and the results of §2.2.

3 Empirical Results

3.1 Simulation study

In this section, we follow Breidt and Opsomer (2000) in the design of a simulation study comparing the performance of several parametric and nonparametric estimators:

REG	parametric regression	Särndal et al. (1992, ch. 6)
PS	poststratification	Cochran (1977, p. 134)
LLR	local linear model-assisted	Breidt and Opsomer (2000)
MB	penalized spline model-based	Dorfman (1992), Zheng and Little (2003)
MA	penalized spline model-assisted	equation (9)

The first two estimators are parametric estimators (corresponding to polynomial and piecewise constant mean functions) and the last three are nonparametric. The LLR estimator is constructed in the same manner as the spline model-assisted estimator in (9), but with $\widehat{m}_i$ obtained by π-weighted local linear regression. The MB estimator is constructed by first computing an unweighted spline regression fit for the sample (say, $\widehat{m}_{i,\mathrm{unw}}$), and then using this fitted model to predict the unobserved elements in the population:

$$\widehat{t}_{y,\mathrm{MB}} = \sum_s y_i + \sum_{U\setminus s} \widehat{m}_{i,\mathrm{unw}}.$$

All these estimators are chosen so that they have approximately the same degrees of freedom (df): the REG estimator is based on a polynomial of degree df-1; the PS estimator is based on a division of the x-range into df equally-sized strata; smoothing constants are chosen so that the traces of the population-level smoothing matrices for LLR and MA are equal to df (see Hastie and Tibshirani, 1990, p.52); and the roughness penalty is chosen so that the trace of an unweighted sample-level smoothing matrix for MB is df. Degrees of freedom for the weighted sample smoothing matrices in LLR, MA, and MB will vary from sample to sample, but are approximately equal to df. Two levels of df are considered: 4 and 7. In the simulation experiments discussed below, we have excluded the classical Horvitz-Thompson and linear regression estimators, as they are based on fewer df and hence not directly comparable. In results not reported here, both estimators

perform poorly relative to the more flexible estimators considered above for most response variables. Of the three nonparametric procedures, two are model-assisted (LLR and MA) and one is model-based (MB). In LLR, we use the Epanechnikov kernel, $0.75(1-t^2)I_{\{|t|\leq 1\}}$. For both spline methods, we use $K = \min\{n/4, 35\} = 25$ knots following equation (5.9) of Ruppert et al. (2003). We consider eight mean functions:

$$
\begin{aligned}
\texttt{linear}: &\quad m_1(x) = 1 + 2(x - 0.5), \\
\texttt{quadratic}: &\quad m_2(x) = 1 + 2(x - 0.5)^2, \\
\texttt{bump}: &\quad m_3(x) = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2), \\
\texttt{jump}: &\quad m_4(x) = \{1 + 2(x - 0.5)I_{\{x\leq 0.65\}}\} + 0.65 I_{\{x>0.65\}}, \\
\texttt{cdf}: &\quad m_5(x) = \Phi\left(\tfrac{1.5-2x}{\sigma}\right), \text{ where } \Phi \text{ is the standard normal cdf}, \\
\texttt{exponential}: &\quad m_6(x) = \exp(-8x), \\
\texttt{cycle1}: &\quad m_7(x) = 2 + \sin(2\pi x), \\
\texttt{cycle4}: &\quad m_8(x) = 2 + \sin(8\pi x),
\end{aligned}
$$

with $x \in [0, 1]$. These represent a range of correct and incorrect model specifications for the various estimators considered. For m_1 and m_2, the models are polynomial; the remaining mean functions represent various departures from the polynomial model. The function m_3 is linear over most of its range, except for a "bump" present for a small portion of the range of x_k. The mean function m_4 is not smooth. The sigmoidal function m_5 is the mean of a binary random variable described below, and m_6 is an exponential curve. The function m_7 is a sinusoid completing one full cycle on $[0, 1]$, while m_8 completes four full cycles. The population x_k are generated as independent and identically distributed (iid) uniform(0,1) random variables. The population values y_{ik} $(i = 1, \ldots, 8)$ are generated from the mean functions by adding iid N(0,σ^2) errors in all cases except `cdf`. The `cdf` response variable consists of binary measurements generated from the linear population via $y_{5k} = I_{\{y_{1k}\leq 1.5\}}$, so that the finite population mean of y_5 is $N^{-1}\sum_{k\in U} I_{\{y_{1k}\leq 1.5\}}$, the

finite population cdf of y_1 at the point $t = 1.5$. We use two possible values for the model standard deviation of the errors: $\sigma = 0.1$ and 0.4. The population is of size $N = 1000$.

Samples are generated from one of two designs of size $n = 50$: simple random sampling without replacement (SI), or stratified simple random sampling without replacement (ST). ST uses seven strata with boundaries given by equally-spaced quantiles of x and with stratum sample sizes (13, 6, 3, 3, 4, 7, 14). Hence, samples drawn via ST are sparse near 0.5 and dense near 0 and 1.

For each combination of mean function, standard deviation, degrees of freedom, and design, 1000 replicate samples are selected and the estimators are calculated. For each sample, a single set of weights corresponding to each estimator (e.g., using (10) for MA weights) is computed and applied to all eight study variables, as would be common practice in applications. As the population is kept fixed during these 1000 replicates, we are able to evaluate the design-averaged performance of the estimators, including the design bias, design variance and design mean squared error. For nearly all cases in this simulation, the relative design biases $(\mathrm{E}_p\widehat{t}_y - t_y)/t_y$ were less than one percent for all estimators, and are not tabled. Table 1 shows the ratios of MSE's for the various estimators to the MSE for the penalized spline model-assisted estimator (MA). The performance of all estimators is increasingly similar as the model variance increases.

[Table 1 about here.]

In most cases, MA is competitive or better than the parametric estimators (MSE ratios ≥ 0.95), though REG with 4 df is better for `exponential` and much better for `cycle1`; a one-cycle sinusoid is fitted extremely well with a cubic polynomial. In each of these cases, however, the penalized spline estimator with more df is at least as good as REG. PS is not bad for `cdf`; the data are mostly one before 0.7 and mostly zero thereafter, so a piecewise

constant fits well. Generally, PS is considerably worse than the other estimators.

The two nonparametric model-assisted estimators MA and LLR are nearly identical, with the exceptions arising in the `bump` and `cycle` variables. In particular, the localized fitting of the kernel-based LLR is better at tracking the multiple curves of the `cycle4` response variable. Under ST, on the other hand, MA dominates LLR for `bump`, since the design in this case generates gaps in the middle of the covariate space, and hence favors the spline methods. Finally, MA and MB are nearly identical under SI, but the model-assisted estimator estimator is sometimes much better, and never much worse, than the model-based estimator under ST sampling.

One common concern when using nonparametric regression techniques is how sensitive the results are to the choice of the smoothing parameter. Clearly, the df (and hence the penalty constant α) has an effect on the MSE of MA, but Table 1 suggests that gains in efficiency over other estimators can be obtained for a variety of choices of df, even when df is chosen by default (without any input from the user or the data). That is, for many reasonable choices of df, the MA estimator is expected to have performance as good or better than the estimators considered here, for a broad range of response variables and study designs. MA can therefore be particularly useful in the context of large-scale survey sampling, in which the same set of regression weights (with a single choice for the smoothing parameter) is often used for a large number of different variables, as was done in the simulation experiment described above.

In any survey application using regression weights, negative weights are clearly undesirable. In the simulation above, there are (2 designs)$\times$(2 df)$\times$(1000 replications)$\times$ (50 weights)$= 200,000$ weights generated for each estimator. Among these, there were 902 negative REG weights, 145 negative LLR weights, and 2 negative MA weights. In practice, negative survey weights are often a result of inappropriate model selection. As

nonparametric models are less likely to result in a severely misspecified population model, they tend to result in less frequent occurrences of negative weights. The superior performance of the MA over the LLR in this respect can be explained by the fact that in regions with very low sampling density, LLR suffers from highly variable fits, which can also occasionally lead to unstable regression weights. As MA is a global regression method, it is less sensitive to the presence of such regions in the data.

A problem shared by all the estimators considered here is the negative finite-sample bias of the design variance estimator defined in Theorem 2.2. This bias is reduced, but not eliminated, for all estimators by the use of the weighted residual variance estimator suggested by Särndal (1982). We found that the coverage of nominal 95% confidence intervals using the weighted residual technique averages about 92% across the various estimators for the sample size of 50. We do not report the details here. Though negative bias of the variance estimator and associated undercoverage are not unique to MA, this problem certainly warrants further investigation.

3.2 Application

In many natural resource surveys, information collected in the field can be supplemented with spatially-referenced data maintained in a geographic information system (GIS). These GIS data are often available at little or no extra cost, especially compared to the cost of collecting the field survey data. To illustrate the applicability of the penalized splines model-assisted approach, we will consider the estimation for a multi-phase natural resource survey: the *Forest Health Monitoring* (FHM) survey in a Central Utah ecoregion.

The purpose of the FHM is to collect information on the ecological and environmental status of public and private forestlands. It is designed as a subsample of the much larger *Forest Inventory and Analysis* (FIA), a field survey conducted by the U.S. Forest Service

(Gillespie, 1999). The FIA is a stratified systematic sample of forest tracts for which a large number of forest and tree characteristics are measured. The FHM data elements are collected at 71 of the 3,107 FIA locations in the region of interest. With such a limited sample size, FHM survey estimates will not be very reliable (this problem would become even more severe if estimates are desired for specific domains within the ecoregion, but we will not further explore domain estimation here). Auxiliary information that can be used in the creation of more efficient FHM estimators is available at two levels: through the FIA, and through a GIS. We will use a grid of 67,216 GIS points covering the ecoregion of interest, and each of the FIA (and FHM) sample locations is matched to its closest GIS point. We will treat the set of 67,216 GIS points as if they were the universe of locations, with FIA locations as a first phase of sampling, and FHM locations as a second phase.

Table 2 defines the relevant variables for the GIS universe and the two levels of sampling. We will only consider estimation of the FHM variable LICHEN here, but it should be noted that in practice, the estimation approach would be used for a large range of other ecological variables collected as part of the FHM survey. The GIS and FIA variables in Table 2 are a small subset of those available at each level, and had been previously identified by Forest Service staff as appropriate predictors of forest health variables.

[Table 2 about here.]

The following two models were fitted using penalized splines regression:

$$\begin{aligned} \mathrm{E}_{\mathrm{GIS}}(\mathrm{LICHEN}) &= m_1(\mathrm{HILLSHD}) + z_{\mathrm{NLCD}}^T \gamma_1 \\ \mathrm{E}_{\mathrm{FIA}}(\mathrm{LICHEN}) &= m_2(\mathrm{CRCOV}) + m_3(\mathrm{AGEMAX}) + m_4(\mathrm{BAMAX}) + z_{\mathrm{NLCD}}^T \gamma_2 \end{aligned}$$

where m_1–m_4 were represented by splines and γ_1, γ_2 are parameters for the NLCD categories in both models. These models were fitted in S-Plus using `gam()` and the `ps()`

penalized spline smoother, which is based on the procedures described in Eilers and Marx (1996). This is equivalent to the semiparametric additive model estimator described in §2.4, except that `ps()` uses B-splines instead of polynomial splines to improve the numerical properties of the algorithm, and `gam()` centers each additive term around its sample mean. For this illustration, the degree of the splines was set to 2, with 15 equally spaced knots and the penalty constant set at $\alpha = 1$ for each smooth function. No attempt was made to "optimize" these values for the individual functions to be fitted. Figures 1 and 2 display the fits obtained for both models.

[Figure 1 about here.]

[Figure 2 about here.]

The fitted values from both models were used in equation (15) to compute the estimate for the average number of lichen species for the region. For comparison, we computed the Horvitz-Thompson estimator (1), and a parametric version of model-assisted estimator (15) which treats each of m_1–m_4 as a linear function in the corresponding variable. We also estimated the standard deviation of the estimators by extending Theorem 2.2, treating both phases of sampling as simple random samples with replacement. This is clearly a simplification, but since the sampling was performed using systematic sampling at both phases, no design-unbiased estimator of the variance is available. Table 3 shows the results. Based on estimated standard errors, the precision of estimates of average number of lichen species present is greatly improved through the use of the auxiliary information. Linear model-assisted estimation decreases the estimated standard error by over 30%, while the semiparametric approach decreases that by an additional 25%.

[Table 3 about here.]

4 Conclusion

Penalized splines are a promising new class of nonparametric regression estimators, and in this paper, we explored their applicability to the model-assisted survey context. As a result of the close connections between penalized splines and linear regression, the survey estimator shares many of the desirable properties of well-known survey estimators such as the GREG, while providing a more flexible model. Because of the latter aspect, penalized splines can improve the efficiency of estimators in situations where linear models are not appropriate.

One of the key strengths of penalized splines is how easy it is to incorporate them into more complicated regression models such as the semiparametric additive models discussed in this paper. This aspect, combined with the flexibility of the model-assisted framework, make this a powerful new tool with which to take advantage of the increasing availability of auxiliary data in surveys.

Acknowledgments

The authors thank three referees and the Editor for helpful comments that greatly improved the readability of the manuscript. For computational assistance, we thank Siobhan Everson-Stewart. Breidt and Opsomer acknowledge support from the National Science Foundation and the U.S. Forest Service, and Claeskens acknowledges partial support from the Belgian Federal Science Policy Office.

References

Aerts, M., G. Claeskens, and M. Wand (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference 103*, 455–470.

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

Breidt, F. J. and J. D. Opsomer (2003). Semiparametric additive models in complex survey estimation. Manuscript.

Cassel, C. M., C. E. Särndal, and J. H. Wretman (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics 12*, 3–32.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 376–382.

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

Gillespie, A. J. R. (1999). Rationale for a national annual forest inventory program. *Journal of Forestry 97*, 16–20.

Hájek, J. (1971). Comment on a paper by D. Basu. In V. P. Godambe and D. A. Sprott (Eds.), *Foundations of Statistical Inference*, pp. 236. Toronto: Holt, Rinehart and Winston.

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

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without re-

placement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

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

Jiang, J. (1996). REML estimation: asymptotic behavior and related topics. *Ann. Statist. 24*(1), 255–286.

Nychka, D. and D. Cummins (1996). Comment on "Flexible smoothing with B-splines and penalties" by P.H.C. Eilers and B.D. Marx. *Statistical Science 89*, 104–105.

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

Rogers, P., D. Atkins, M. Frank, and D. Parker (2001, May). Forest Health Monitoring in the Interior West: A baseline summary of forest health issues, 1996-1999. General Technical Report RMRS-GTR-75, USDA Forest Service, Rocky Mountain Research Station.

Ruppert, D. and R. J. Carroll (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics 42*, 205–223.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference 7*, 155–170.

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

Searle, S. R., G. Casella, and C. E. McCulloch (1992). *Variance components*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics.

New York: John Wiley & Sons Inc. A Wiley-Interscience Publication.

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

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman and Hall.

Wu, C. (1999). *The effective use of complete auxiliary information from survey data*. Ph. D. thesis, Simon Fraser University, Vancouver, BC, Canada.

Zheng, H. and R. J. A. Little (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics 19*, 99–117.

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.

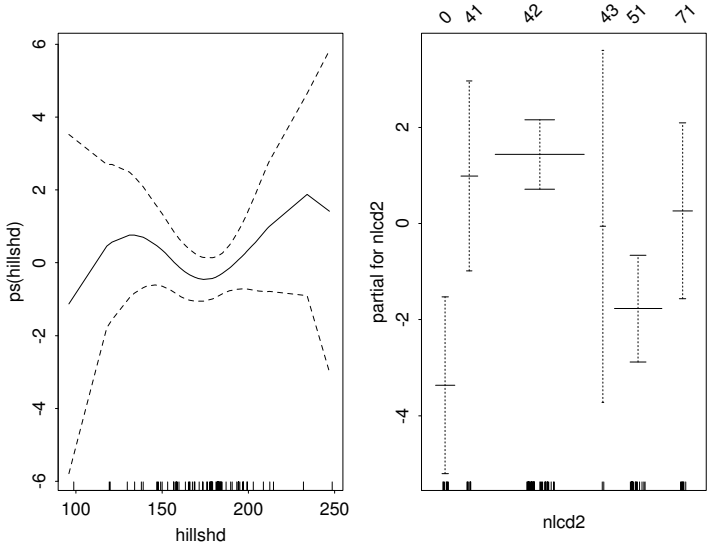

Figure 1: Estimated mean functions and 95% pointwise confidence bands for GIS-level model for variable LICHEN. Left plot is for the estimated function $\widehat{m}_1$, right plot shows the estimated parameter values $\widehat{\gamma}_1$.

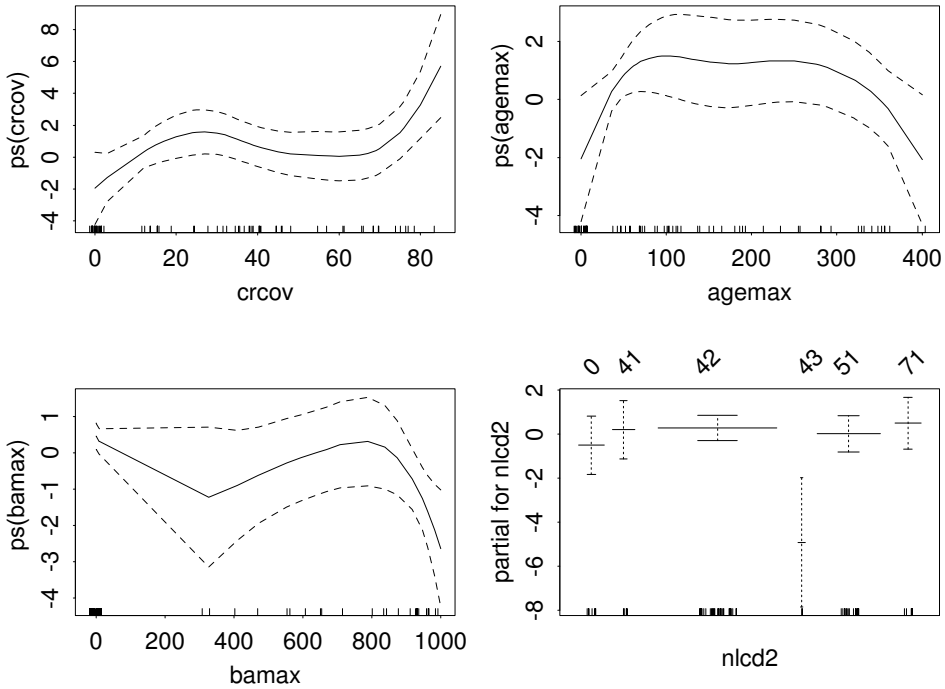

Figure 2: Estimated mean functions and 95% pointwise confidence bands for FIA-level model for variable LICHEN. Three line plots are for the estimated functions $\widehat{m}_2$, $\widehat{m}_3$ and $\widehat{m}_4$, last plot shows the estimated parameter values $\widehat{\gamma}_2$.

			REG		PS		LLR		MB	
Response	σ	df	SI	ST	SI	ST	SI	ST	SI	ST
	0.1	4	1.02	0.99	2.78	2.48	1.00	0.99	1.00	0.91
line	0.1	7	1.12	1.00	1.74	1.55	1.06	1.01	1.00	0.96
	0.4	4	1.02	0.99	2.03	1.85	1.00	0.99	1.00	0.91
	0.4	7	1.12	1.00	1.46	1.31	1.06	1.01	1.00	0.96
	0.1	4	0.99	1.00	1.56	1.29	1.01	1.00	0.99	1.08
quadratic	0.1	7	1.12	1.00	1.22	1.08	1.05	1.02	1.00	0.96
	0.4	4	1.01	1.00	1.33	1.19	1.01	1.00	0.99	1.02
	0.4	7	1.12	1.00	1.14	1.03	1.05	1.02	1.00	0.96
	0.1	4	1.35	1.14	1.77	2.33	1.06	1.07	0.99	1.41
bump	0.1	7	3.00	1.23	1.03	1.66	0.95	1.13	0.99	0.98
	0.4	4	1.30	1.09	1.67	1.94	1.05	1.05	0.99	1.23
	0.4	7	2.59	1.16	1.05	1.44	0.97	1.10	0.99	0.97
	0.1	4	1.30	1.09	1.46	1.50	1.07	1.03	0.99	1.24
jump	0.1	7	1.68	1.11	1.50	1.43	1.04	1.00	0.99	0.96
	0.4	4	1.22	1.07	1.35	1.37	1.05	1.03	0.99	1.13
	0.4	7	1.53	1.09	1.38	1.30	1.05	1.00	0.99	0.96
	0.1	4	1.20	1.03	1.04	1.29	0.97	0.98	0.99	1.65
cdf	0.1	7	1.66	1.04	1.15	1.11	0.97	0.97	0.99	1.05
	0.4	4	1.21	1.03	1.05	1.31	0.97	0.99	0.99	1.63
	0.4	7	1.70	1.05	1.16	1.12	0.97	0.97	0.99	1.04
	0.1	4	0.86	0.95	1.96	1.43	0.93	0.98	0.99	0.90
exponential	0.1	7	1.07	0.99	1.60	1.23	1.03	1.02	0.99	0.95
	0.4	4	0.91	0.97	1.64	1.28	0.95	0.99	0.99	0.90
	0.4	7	1.09	1.00	1.37	1.12	1.04	1.02	1.00	0.95
	0.1	4	0.31	0.84	2.09	5.31	0.94	0.97	0.95	1.23
cycle1	0.1	7	1.02	1.00	3.76	3.76	1.25	1.09	0.99	0.95
	0.4	4	0.41	0.90	1.97	3.95	0.96	0.99	0.96	1.10
	0.4	7	1.06	1.00	2.71	2.62	1.18	1.06	0.99	0.95
	0.1	4	1.00	1.02	1.10	1.06	1.03	1.04	0.99	0.89
cycle4	0.1	7	2.30	1.38	1.32	1.18	0.86	0.76	0.99	1.06
	0.4	4	1.00	1.01	1.10	1.06	1.03	1.04	0.99	0.89
	0.4	7	2.28	1.37	1.32	1.18	0.86	0.77	0.99	1.05

Table 1: Mean square error ratios greater than one favor the penalized spline model-assisted estimator (MA). Based on 1000 replicate samples (SI=simple random sampling without replacement, ST= stratified simple random sampling with seven strata) of size $n = 50$ from a fixed population of size $N = 1000$. All five estimators use the same degrees of freedom (df) for fitting. Penalized spline estimators use 25 knots and truncated linear basis. Local linear regression estimators are computed with an Epanechnikov kernel and bandwidth $h = 0.2943$ when df= 4 and $h = 0.135$ when df= 7.

Phase	Variable	Description
GIS	HILLSHD	hillshade (index of solar radiation)
	NLCD	vegetation cover type (6 categories)
FIA	CRCOV	tree crown cover (%)
	AGEMAX	max tree age (years)
	BAMAX	max tree basal area (sq. in.)
FHM	LICHEN	lichen species present (count)

Table 2: Variables used in the construction of the Forest Health Monitoring estimator.

	Horvitz-Thompson	Linear MA	Semiparametric MA
Estimate	3.62	2.92	2.67
Standard Error	0.36	0.25	0.16
		(69%)	(44%)

Table 3: Estimates of region average for variable LICHEN, using three estimation methods, with corresponding estimated standard deviations. Numbers in parentheses are the ratio of the estimated standard deviations relative to that of the Horvitz-Thompson estimator.