Model-Assisted Estimation of Forest Resources With Generalized Additive Models

Jean D. OPSOMER, F. Jay BREIDT, Gretchen G. MOISEN, and Göran KAUERMANN

Multiphase surveys are often conducted in forest inventories, with the goal of estimating forested area and tree characteristics over large regions. This article describes how design-based estimation of such quantities, based on information gathered during ground visits of sampled plots, can be made more precise by incorporating auxiliary information available from remote sensing. The relationship between the ground visit measurements and the remote sensing variables is modeled using generalized additive models. Nonparametric estimators for these models are discussed and applied to forest data collected in the mountains of northern Utah. Model-assisted estimators that use the nonparametric regression fits are proposed for these data. The design context of this study is two-phase systematic sampling from a spatial continuum, under which properties of model-assisted estimators are derived. Difficulties with the standard variance estimation approach, which assumes simple random sampling in each phase, are described. An alternative assessment of estimator performance based on a synthetic population is implemented and shows that using the model predictions in a model-assisted survey estimation procedure results in substantial efficiency improvements over current estimation approaches.

KEY WORDS: Calibration; Multiphase survey estimation; Nonparametric regression; Systematic sampling; Variance estimation.

1. INTRODUCTION

Accurate estimation of forest resources over large geographic areas is of significant interest to forest managers and forestry scientists. Nationwide forest surveys of the U.S. are conducted by the U.S. Department of Agriculture Forest Service Forest Inventory and Analysis (FIA) program (U.S. Department of Agriculture Forest Service 1992; Frayer and Furnival 1999; Gillespie 1999). In these surveys, design-based estimates of quantities like total tree volume, growth and mortality, or area by forest type are produced on a regular basis. We consider the estimation of such quantities within a 2.5 million-ha ecological province (Bailey, Avers, King, and McNab 1994) that includes the Wasatch and Uinta Mountain Ranges of northern Utah. Forests in the area consist of pinyon-juniper, oak, and maple generally in the lower elevations and lodgepole pine, ponderosa pine, aspen, and spruce-fir generally in the higher elevations. Many forest types intermix and swap elevation zones according to other topographic variables, such as aspect and slope. Besides having ecological diversity, the area hosts numerous large ownerships, including national forests, Indian reservations, national parks and monuments, state land holdings, and private lands. Each owner group faces different land management issues requiring precise forest resource information. Figure 1 displays the region of interest and the sample points collected in the early 1990s for the survey that we consider here. Although this article focuses on this particular example, the general approach proposed here is applicable to other natural resource estimation problems as well.

Currently, forest survey data are collected through a two-phase systematic sampling procedure. In phase one, remote sensing data and geographical information system (GIS) coverage information are extracted on an intensive sample grid. Phase two consists of a field-visited subset of the phase one grid. During these field visits, several hundred variables are collected, ranging from individual tree characteristics and size measurements to complex ratings on scales of ecological health.

Once the data are collected, estimates of population totals and related quantities need to be calculated and tabulated for the overall region, as well as for various domains defined by political subdivisions, types of forest, ownership category, and other factors. There are literally thousands of estimates in the core tables put out by the FIA, with an even larger number of potential "custom estimates" that can be requested by data users. It is desirable that these estimates be internally consistent, in the sense that tables "add up"; that is, the estimate of a sum of subdomain totals equals the sum of the subdomain total estimates. Herein, the problem of making sensible estimates for a large number of quantities in a straightforward and internally consistent way is called *generic inference*. This can be contrasted with *specific inference*, in which the statistician responsible for producing estimates is studying a small number of variables and is able to build custom models for the dataset at hand.

In the generic inference context, the statistician has neither time nor resources to conduct detailed analyses of all response variables. Therefore, often the only practical way to produce estimates is through design-based estimation, in which survey weights are constructed and applied to all variables and domains of interest. These weights are derived from the sampling design but are adjusted based on ancillary information available for the sampled universe and/or collected as part of the survey. The ancillary information is used to calibrate the survey weights (making them sum to population quantities that are known or precisely estimated) and to improve the efficiency of the survey estimators. Once the weights are computed, users of the data can easily produce estimates for any variable of interest. Subdomain analyses are also simplified, because the linear form of the estimators guarantees internal consistency.

A large number of techniques for adjusting survey weights based on auxiliary information are available. The use of auxiliary information in surveys dates back at least to Laplace (see Cochran 1978), who used a ratio estimator. The earliest references to regression in surveys include those by Jessen

Jean D. Opsomer is Professor, Department of Statistics, Iowa State University, Ames, IA 50011 (E-mail: *jopsomer@iastate.edu*). F. Jay Breidt is Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523. Gretchen G. Moisen is Research Forester, U.S. Department of Agriculture Forest Service, Rocky Mountain Research Station, Ogden, UT 84401. Göran Kauermann is Professor, Department of Economics, University of Bielefeld, 33501 Bielefeld, Germany. This work was supported in part by the U.S. Department of Agriculture Forest Service Rocky Mountain Research Station grants RJVA 02-JV-11222007-004 and 01-JV-11222007-307 and National Science Foundation grants DMS-02-04531 and DMS-02-04642.

Journal of the American Statistical Association
June 2007, Vol. 102, No. 478, Applications and Case Studies
DOI 10.1198/016214506000001491

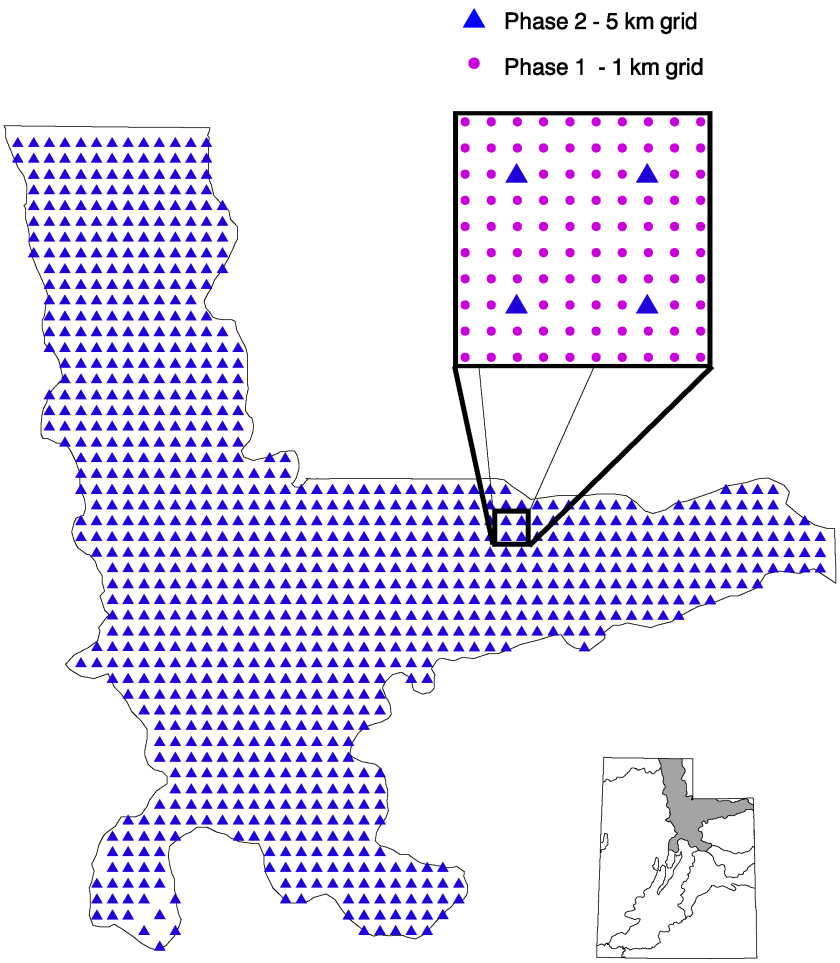

Figure 1. Representation of the Study Region in Northern Utah. Each triangle represents a field-visited phase two plot. Each dot in the magnified section represents a remotely sensed phase one plot. In the notation of the text, there are $n_1 = 24{,}980$ phase one plots, located $\delta_1 = \delta_2 = 1$ km apart in both horizontal and vertical dimensions, and $n_2 = 968$ phase two plots, located $h_1\delta_1 = h_2\delta_2 = 5$ km apart in both dimensions.

(1942) and Cochran (1942). Typically, auxiliary information is incorporated into the survey inference through parametric linear models, leading to the familiar ratio and regression estimators (e.g., Cochran 1977), poststratification estimators (Holt and Smith 1979), best linear unbiased estimators (Brewer 1963; Royall 1970), generalized regression estimators (Cassel, Särndal, and Wretman 1977; Särndal 1980; Robinson and Särndal 1983), and related estimators (Wright 1983; Isaki and Fuller 1982). Fuller (2002) has provided an excellent review. Recent advances in the use of auxiliary information include nonlinear estimation (Wu and Sitter 2001), nonparametric survey regression estimation (Kuo 1988; Dorfman 1992; Dorfman and Hall 1993; Chambers, Dorfman, and Wehrly 1993; Breidt and Opsomer 2000), and the calibration point of view (Deville and Särndal 1992).

The approach currently used for the FIA is based on two-phase poststratification (Scott et al. 2004), which does not take advantage of the increasing availability of various inexpensive auxiliary data derived from remote sensing sources. Thus there is a tremendous opportunity to both reduce costs and improve the precision of forest survey estimates. The need for this is all the more pressing because scientists within the Forest Service and other institutions have been using remote sensing and other GIS data to develop predictive and analytical models describing forest characteristics, often with the help of nonparametric or semiparametric techniques. This has been done in the specific inference context, in which significant effort is directed toward finding appropriate models for a small number of important variables. Although these modeling efforts have led to improved understanding of the relationships between key forestry variables and remotely sensed information, so far this has not been reflected in corresponding improvements in forest survey estimates.

This article aims to explain how the results from specific inferential efforts by forestry specialists can be used to improve the quality of their generic inference outputs as well. We do this by incorporating multidimensional nonparametric superpopulation models into the framework of *model-assisted estimation* (Särndal, Swensson, and Wretman 1992) and then applying the resulting estimation approach to the two-phase FIA data. This extends the nonparametric model-assisted methodology of Breidt and Opsomer (2000), who considered only univariate models and single-phase estimation.

The remainder of the article is organized as follows. Section 2 begins by developing the methodological framework needed for this application, including the two-phase systematic sampling design for a continuous spatial domain of in-

terest (Sec. 2.1), two-phase model-assisted survey estimation (Sec. 2.2), and nonparametric model-assisted estimation (Sec. 2.3). We discuss the data and generalized additive model (GAM) fits for the northern Utah mountains forest inventory in Section 3.1, and show how to incorporate the fits into a generic model-assisted estimation strategy in Section 3.2. Evaluation of the efficiency gains from this procedure is complicated by the lack of good variance estimators for systematic sampling, as discussed in Section 4.1, so we evaluate the methodology in a synthetic population in Section 4.2. A brief discussion follows in Section 5.

2. METHODOLOGY

2.1 Two-Phase Systematic Sampling From a Spatial Domain

The northern Utah mountains data were collected in two phases on a regularly spaced grid (see Fig. 1). We describe the design properties of model-assisted estimation in this context by first considering a rectangular spatial domain, $D = [0, L_1] \times [0, L_2]$, where $L_k = n_{1k}\delta_k = n_{2k}h_k\delta_k$, with n_{jk} and h_k as positive integers and δ_k as positive real numbers. On D, we define two grids: a fine phase-one grid and a coarser phase-two subgrid. In dimension k, δ_k denotes the "grid spacing" (in km) for the phase-one grid, and h_k denotes the number of phase-one grid points between phase-two grid points, so that $h_k\delta_k$ represents the grid spacing (in km) for the phase-two grid. In the sampling situation depicted in Figure 1, $\delta_k = 1$ km is the spacing between two neighboring phase-one dots in both the horizontal and vertical dimensions, and $h_k\delta_k = 5$ km is the corresponding spacing between two neighboring phase-two triangles in both dimensions. Then $n_1 = n_{11}n_{12}$ is the phase-one sample size and $n_2 = n_{21}n_{22}$ is the phase-two sample size, both over D. An irregular spatial domain like that in Figure 1 is handled by intersecting it with the rectangle D. We continue to use n_1 and n_2 to denote the sample sizes for both phases over the irregular domain as well.

Implementation of the two-phase systematic sampling design requires random placement of the phase-one grid on D, followed by random selection of the phase-two subgrid. Let u_k represent independent uniform$(0, 1)$ random variables and let d_k represent independent discrete uniform$\{1, 2, \ldots, h_k\}$ random variables, with the u_k and d_k independent of each other. Given $u = (u_1, u_2)$, the phase-one sample is the randomly located grid

$$G_1(u) = \big\{\big((u_1 + i_1 - 1)\delta_1, (u_2 + i_2 - 1)\delta_2\big)\big\} \quad \text{for } i = (i_1, i_2) \in \{1, \ldots, n_{11}\} \times \{1, \ldots, n_{12}\}.$$

Given $d = (d_1, d_2)$, the phase-two sample is the randomly selected subgrid on $G_1(u)$,

$$G_2(u, d) = \big\{\big((u_1 + d_1 + (j_1 - 1)h_1 - 1)\delta_1, (u_2 + d_2 + (j_2 - 1)h_2 - 1)\delta_2\big)\big\} \quad \text{for } j = (j_1, j_2) \in \{1, \ldots, n_{21}\} \times \{1, \ldots, n_{22}\}.$$

The union of all possible phase-two subgrids is the phase-one grid: $\bigcup_d G_2(u, d) = G_1(u)$. Figure 1 shows the realization of $G_1(u)$ (dots) and $G_2(u, d)$ (triangles) for the region of interest in northern Utah. For this dataset, there are exactly $h = h_1h_2 = 25$ possible phase-two samples for each realized phase-one sample.

2.2 Two-Phase Model-Assisted Estimation

To motivate the model-assisted approach, we begin with a discussion of the two-phase difference estimator. Let $z(s)$ denote the response variable of interest, defined for $s \in D$ but observed only for $s \in G_2(u, d)$, and let $z^0(s)$ represent a different variable that is known for all $s \in G_1(u)$. Note that neither $z(\cdot)$ nor $z^0(\cdot)$ is assumed stochastic, and in particular neither depends on the random vectors u and d. Define a set of rectangles, $D_i = [(i_1 - 1)\delta_1, i_1\delta_1] \times [(i_2 - 1)\delta_2, i_2\delta_2]$, which partition the domain D. Then the population total,

$$\theta := \int_D z(v)\,dv = \sum_i \int_{D_i} z(v)\,dv = \int_{[0,1]\times[0,1]} \sum_{s\in G_1(u)} \frac{z(s)}{1/(\delta_1\delta_2)}\,du, \tag{1}$$

can be estimated with the two-phase difference estimator,

$$\hat\theta := \sum_{s\in G_1(u)} \frac{z^0(s)}{1/(\delta_1\delta_2)} + \sum_{s\in G_2(u,d)} \frac{z(s) - z^0(s)}{1/(\delta_1\delta_2 h)} = \sum_{d'} \sum_{s\in G_2(u,d')} \left\{ \frac{z^0(s)}{1/(\delta_1\delta_2)} + \frac{z(s) - z^0(s)}{1/(\delta_1\delta_2)} \frac{I_{\{d=d'\}}}{1/h} \right\}, \tag{2}$$

with $I_{\{d=d'\}} = 1$ if $d = d'$ and 0 otherwise, where the summation over d' is over all possible values for the random pair (d_1, d_2). This is the continuous-domain equivalent of the two-phase difference estimator described by Särndal et al. (1992, p. 358). To simplify notation, we suppress the dependence of $\hat\theta$ and subsequent population total estimators on u and d.

Because the indicator $I_{\{d=d'\}}$ has expectation $1/h$, we have

$$\mathrm{E}(\hat\theta \mid u) = \sum_{d'} \sum_{s\in G_2(u,d')} \frac{z(s)}{1/(\delta_1\delta_2)} = \sum_{s\in G_1(u)} \frac{z(s)}{1/(\delta_1\delta_2)}. \tag{3}$$

Using (1), it is then immediate that

$$\mathrm{E}(\hat\theta) = \int_{[0,1]\times[0,1]} \mathrm{E}(\hat\theta \mid u)\,du = \theta. \tag{4}$$

In addition, by standard results on systematic sampling from a finite population, we have that

$$\mathrm{var}(\hat\theta \mid u) = \frac{|D|^2}{n_2^2}\left(1 - \frac{1}{h}\right) S^2(u), \tag{5}$$

where $|D| = \int_D dv$,

$$S^2(u) = \frac{\sum_d t_d^2(u) - (\sum_d t_d(u))^2/h}{h - 1},$$

and

$$t_d(u) = \sum_{s\in G_2(u,d)} (z(s) - z^0(s)). \tag{6}$$

As shown by (4), the estimator $\hat\theta$ is design-unbiased regardless of the relationship between z and z^0. Its design variance is

given by

$$\begin{aligned}\operatorname{var}(\hat{\theta}) &= \operatorname{var}(\mathrm{E}(\hat{\theta} \mid u)) + \mathrm{E}(\operatorname{var}(\hat{\theta} \mid u)) \\ &= \int_{[0,1]\times[0,1]} (\mathrm{E}(\hat{\theta} \mid u) - \theta)^2 \, du \\ &\quad + \frac{|D|^2}{n_2^2}\left(1 - \frac{1}{h}\right)\mathrm{E}(S^2(u)). \end{aligned} \tag{7}$$

The first component of the variance does not depend on the choice of z^0, but the second component of the variance will be small if z^0 is a good predictor of z, because it depends on "residuals" of the form (6). In the following result, (4) and (7) are combined to show that $\hat{\theta}$ is design-consistent under an asymptotic formulation in which the sampling density in D increases ("infill asymptotics"), assuming integrability conditions on z and z^0. This result is similar to consistency results obtained in design-based stereology (Arnau and Cruz-Orive 1996), but the two-phase structure is novel.

Result 1. If $z(\cdot)$ and $z^0(\cdot)$ are bounded and continuous almost everywhere on D, then $\hat{\theta}$ converges in mean square to θ as $n_{jk} \to \infty$ with D fixed.

In the absence of useful information from the first-phase sample, the simple expansion estimator

$$\hat{\theta}_{\mathrm{exp}} = \sum_{s\in G_2(u,d)} \frac{z(s)}{1/(\delta_1\delta_2 h)}, \tag{8}$$

obtained from (2) with $z^0 \equiv 0$, can be used. In most cases of two-phase sampling, however, relatively inexpensive auxiliary information, $\{\mathbf{X}(s)\}_{s\in G_1(u)}$, is collected at each phase-one site. This information can be used to construct predictors of z guided by a superpopulation model,

$$\mathrm{E}[z(s) \mid \mathbf{X}(s)] = \mu(\mathbf{X}(s)). \tag{9}$$

Typically, $\mu(\cdot)$ is estimated from regression of $\{z(s)\}$ on $\{\mathbf{X}(s)\}$ for $s \in G_2(u,d)$, and the resulting fits $\{\hat{\mu}(\mathbf{X}(s))\}$ for phase one are then substituted into (2) to form the model-assisted estimator

$$\hat{\theta}_{\mathrm{ma}} = \sum_{d'} \sum_{s\in G_2(u,d')} \left\{ \frac{\hat{\mu}(\mathbf{X}(s))}{1/(\delta_1\delta_2)} + \frac{z(s) - \hat{\mu}(\mathbf{X}(s))}{1/(\delta_1\delta_2)} \frac{I_{\{d=d'\}}}{1/h} \right\}. \tag{10}$$

Unlike $z^0(\cdot)$, $\hat{\mu}(\cdot)$ usually does depend on u and d, so the unbiasedness argument in (4) and the variance expression in (7) no longer hold exactly. However, under mild conditions that we do not explore further here (see Kim 2004 for the univariate nonparametric regression case), the model-assisted two-phase estimator should follow the traditional model-assisted paradigm and remain asymptotically design-unbiased and consistent, with approximate variance given by

$$\begin{aligned}\operatorname{var}(\hat{\theta}_{\mathrm{ma}}) &= \int_{[0,1]\times[0,1]} (\mathrm{E}(\hat{\theta}_{\mathrm{ma}} \mid u) - \theta)^2 \, du \\ &\quad + \frac{|D|^2}{n_2^2}\left(1 - \frac{1}{h}\right)\mathrm{E}(S_e^2(u)), \end{aligned} \tag{11}$$

where

$$S_e^2(u) = \frac{\sum_d \hat{t}_d^2(u) - (\sum_d \hat{t}_d(u))^2/h}{h-1},$$

$$\hat{t}_d(u) = \sum_{s\in G_2(u,d)} \big(z(s) - \tilde{\mu}(\mathbf{X}(s))\big),$$

and $\tilde{\mu}(\cdot)$ is obtained from the (hypothetical) regression of $\{z(s)\}$ on $\{\mathbf{X}(s)\}$ for $s \in G_1(u)$. Expression (11) is the extension of the linear model-assisted result of Särndal et al. (1992, p. 362) to a continuously defined population and a nonparametric superpopulation model.

It is now clear why a model can improve the efficiency of the estimator. If the model fits the data well, then the variance of the residuals $z(s) - \tilde{\mu}(\mathbf{X}(s))$ can be expected to be smaller than the variance of the $z(s)$. If the model is misspecified, then the residual variance can be equally large or even (in some extreme cases) larger than the response variable's variance. Hence the efficiency gains of the model-assisted estimator depend on selection of a good model for $\mu(\cdot)$ in (9). However, regardless of the correctness of the model, design consistency is maintained. This characteristic of model-assisted estimation has been established in other contexts and stands in contrast to purely model-based estimation, for which model misspecification can lead to biased or inconsistent estimators. This is an important consideration for generic inference, because any assumed model is unlikely to be equally appropriate across all of the variables for which estimates need to be constructed.

The estimator $\hat{\theta}_{\mathrm{ma}}$ has some additional desirable properties if the regression method used to obtain $\hat{\mu}(\cdot)$ is linear, in the sense that $\hat{\mu}(\mathbf{X}(v)) = \sum_{s\in G_2(u,d)} r(v,s)z(s)$ for a set of regression weights $\{r(v,s)\}$ that depend on the auxiliary variables $\{\mathbf{X}(s)\}_{s\in G_1(u)}$, but not on the $\{z(s)\}_{s\in G_2(u,d)}$. If $\hat{\mu}(\cdot)$ is linear, then

$$\begin{aligned}\hat{\theta}_{\mathrm{ma}} &= \sum_{s\in G_2(u,d)} \left\{ \sum_{v\in G_1(u)} \frac{r(v,s)}{1/(\delta_1\delta_2)} - \sum_{v\in G_2(u,d)} \frac{r(v,s)}{1/(\delta_1\delta_2 h)} \right. \\ &\qquad \left. + \frac{1}{1/(\delta_1\delta_2 h)} \right\} z(s) \\ &= \sum_{s\in G_2(u,d)} w(s)z(s), \end{aligned} \tag{12}$$

with survey weights $\{w(s)\}$ independent of the $\{z(s)\}$. This survey-weighted form holds for generalized regression estimators, including ratio and linear regression estimators, as well as the poststratification estimator currently used for the FIA. The survey weights are ideal for generic inference, because they can be used for any variables collected in the same survey, and to the extent that such variables follow model (9), they will benefit from the efficiency gain. Thus it is desirable to specify model (9) as flexibly as possible.

2.3 Model-Assisted Estimation Using Generalized Additive Models

Suppose now that $\mu(\mathbf{X}(s))$ is the GAM

$$\begin{aligned}\mu(\mathbf{X}(s)) &= \mathrm{E}(z(s) \mid \mathbf{X}(s)) \\ &= g\big(m_1(\mathbf{X}_1(s)) + \cdots + m_r(\mathbf{X}_r(s))\big) \end{aligned} \tag{13}$$

for some known link function $g(\cdot)$ and unknown smooth functions $m_k(\cdot)$, $k = 1, \ldots, r$, where the $\mathbf{X}_k(s)$ are known subsets

of the vector $\mathbf{X}(s)$. Such a model can be fitted using, for instance, the `gam()` local scoring estimation routines (Hastie and Tibshirani 1990) in S–PLUS. Given a set of estimated functions $\hat{m}_k(\cdot), k = 1, \ldots, r$, model predictions $\hat{\mu}(\mathbf{X}(s)) = g(\hat{m}_1(\mathbf{X}_1(s)) + \cdots + \hat{m}_r(\mathbf{X}_r(s)))$ are then readily calculated for all phase-one points.

When the link function $g(\cdot)$ is the identity link, model (13) is referred to as an additive model, and the resulting regression estimator $\hat{\mu}(\cdot)$ is linear in the sense described in Section 2.2. But if $g(\cdot)$ is not the identity link, then local scoring estimators are not linear, and the resulting estimator $\hat{\theta}_{\text{gam}}$ is no longer a linear combination of the $\{z(s)\}_{s\in G_2(u,d)}$, so that weights are not available. In the next section we discuss an approach for obtaining weights from a GAM for the forestry application.

3. APPLICATION TO FOREST INVENTORY

3.1 Generalized Additive Models for the Forest Inventory Data

Field data used in this study were collected on a 5-km sample grid (Fig. 1). On the $n_2 = 968$ phase-two sample plots, numerous forest site variables and individual tree measurements were collected, including a binary classification (`FOREST`) of the plot into "forest" or "nonforest." The `FOREST` variable is critical in the inventory because many other response variables are defined to be zero on nonforested sites. In this article we consider five additional FIA variables, all of which follow this definitional constraint: `NVOLTOT`, total wood volume in cubic ft per acre; `BA`, tree basal area per acre; `BIOMASS`, total wood biomass in tons per acre; `CRCOV`, percent crown cover; and `QMDALL`, quadratic mean diameter in inches. For the purpose of illustration, we develop model-assisted estimators for these six variables; however, it should be noted that in practice, our estimation method would be applied generically to all other variables collected as part of the FIA.

The phase-one data for this study consist of remotely sensed information extracted on a 1-km grid ($n_1 = 24{,}980$ points) in which the 5-km grid of field plots is embedded (see Fig. 1). The ancillary variables used in our models came from three sources:

1. Digital elevation models produced by the U.S. Defense Mapping Agency, which provided elevation (`ELEV90CU`), transformed aspect (`TRASP90`), and slope (`SLP90CU`).
2. 30-m resolution thematic mapper (TM) imagery, from which we extracted the vegetation cover type of the U.S. National Land Cover dataset (Vogelmann et al. 2001) collapsed to seven vegetation classes (`NLCD7`). In addition, letting `MRLC00B`k denote the kth TM spectral band, we used `MRLC00B5` by itself and we computed a Normalized Difference Vegetation Index (`NDVI`) as (`MRLC00B4` − `MRLC00B3`)/(`MRLC00B4` + `MRLC00B3`).
3. Spatial coordinates (`Xs` and `Ys`).

More details on these variables have been given by Moisen and Edwards (1999) and Frescino, Edwards, and Moisen (2001). Those works (as well as Moisen and Frescino 2002) illustrate the use of parametric and nonparametric models relating remotely sensed data to forest attributes observed during field visits. Taking a similar approach here, we model the response variable `FOREST` as a nonparametric function of the ancillary predictor variables mentioned earlier through a GAM with a logit link function $g(\cdot)$. We fitted model (13) using `gam()` in S–PLUS. Component functions were obtained through *loess* smoothers with local polynomials of degree 1 and a relatively large smoothing parameter. (See Opsomer 2002 for an explanation of the loess smoothing method and smoothing parameter selection.) Values for the smoothing parameters were selected through trial and error to achieve a visually reasonable fit while trying to avoid a model with excessive degrees of freedom. Predictor variables `ELEV90CU`, `TRASP90`, `SLP90CU`, `MRLC00B5`, and `NDVI` entered the model as univariate smooth terms, whereas `Xs` and `Ys` contributed as a bivariate smooth function and `NLCD7` entered as a categorical variable in the model. Plots of the smooth terms in the `FOREST` model are shown in Figure 2.

3.2 Model-Assisted Estimation Using the Generalized Additive Model Fits

To allow incorporation of the result of the GAM fitting in the generic estimation for FIA survey variables, survey weights must be constructed. Because of the presence of the logit link in the model for `FOREST`, the regression estimator is not linear (as noted in Sec. 2.2), so that no regression weights are available to construct linear survey weights. One possible solution to this problem is to use a "model calibration" step as done by Wu and Sitter (2001) in single-phase estimation for nonlinear regression models. This approach uses the GAM predictions for the `FOREST` variable on phase one, which we denote here by $\hat{\mu}_F(s)$, as an auxiliary variable in a linear regression model,

$$\begin{aligned}[z(s)]_{s\in G_2(u,d)} &= [\hat{\mu}_F(s)\mathbf{X}'(s)]_{s\in G_2(u,d)}\boldsymbol{\beta} + \text{error} \\ &= \mathbf{C}\boldsymbol{\beta} + \text{error}, \end{aligned} \tag{14}$$

so that the row vector of n_2 regression weights is given by $[r(v,s)] = [\hat{\mu}_F(v)\mathbf{X}'(v)](\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'$. Survey weights then can be constructed using (12) and applied to all survey variables. Because of the dependence of the regression weights on the $\hat{\mu}_F(s)$, the resulting model-assisted estimator is not strictly linear, but can be considered approximately linear by treating the $\hat{\mu}_F(s)$ as fixed with respect to the design. (Results in Wu and Sitter 2001 suggest that this approximation is legitimate asymptotically.)

Although this is a suitable approach in general, the special structure of the relationship between the presence/absence of forest and the other variables suggests a more appropriate calibration regression model than that given in (14). For every phase-one site $s \in G_1(u)$, we construct an indicator variable that is 1 when the GAM-predicted probability of forest is greater than the empirical proportion of forest in phase two,

$$\hat{I}_F(s) = \begin{cases} 1 & \text{if } \hat{\mu}_F(s) \geq \hat{\theta}_{F,\text{exp}}/|D| \\ 0 & \text{otherwise,} \end{cases}$$

with $\hat{\theta}_{F,\text{exp}}$ as in (8) with z taken to be the binary variable `FOREST`. Then the regression model (14) is replaced by

$$[z(s)]_{s\in G_2(u,d)} = [\mathbf{X}'(s) \times \hat{I}_F(s)]_{s\in G_2(u,d)}\boldsymbol{\beta} + \text{error}, \tag{15}$$

so that the covariates consist entirely of interactions between the forest indicator and other auxiliary variables. The variables that we used in $\mathbf{X}(s)$ are linear terms for the same variables as

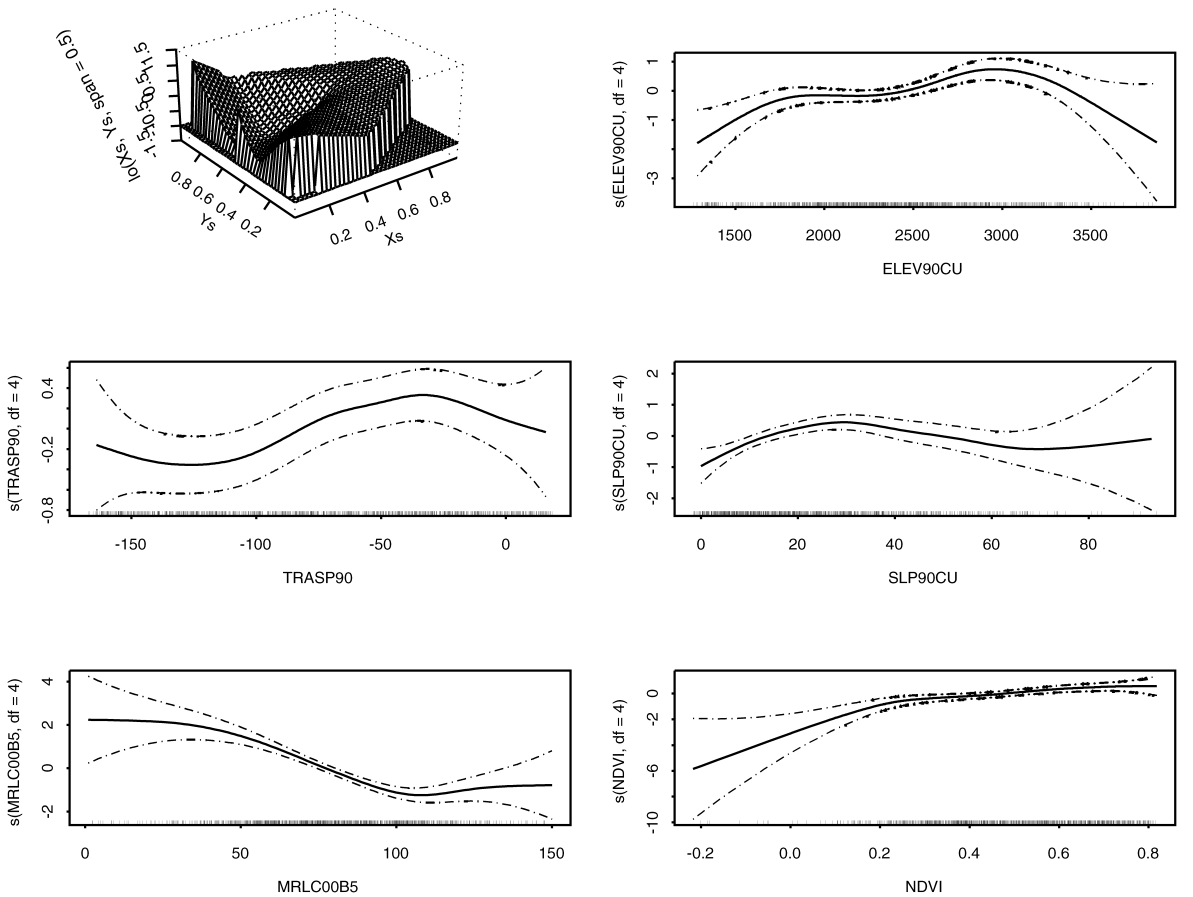

Figure 2. GAM Model Fits for Binary Indicator of Forest/Nonforest (FOREST).

used in the GAM FOREST model, with nonforest categories in NLCD7 collapsed to ensure full rank. Note that this regression model predicts 0 for the response variable at any phase-one site for which FOREST is predicted to be 0 by $\hat{I}_F(s)$.

Regression weights and the model-assisted estimator $\hat{\theta}_{\text{ma}}$ are built from this regression model in the same manner as described for model (14). Once again, the resulting model-assisted estimator is not strictly linear, but can be approximated as linear. This approximation yields a survey weighted form for the estimator $\hat{\theta}_{\text{ma}}$, with weights that can be applied to all survey variables as required for generic inference. When these survey weights are applied to the variable FOREST, the estimator is very close, but not identical, to that obtained directly from the GAM.

The weighted estimator is internally consistent when applied to domains within the region of interest; for example, estimated BIOMASS for the entire region is the sum of the estimated BIOMASS for each ownership domain within the region (national forest, Indian reservation, etc.). Furthermore, the survey weights are calibrated for the auxiliary variables on the part of phase one predicted to be forest; that is, the phase-two estimated totals agree exactly with the phase-one estimated totals on predicted forest: $\sum_{s\in G_2(u,d)} w(s)(\mathbf{X}'(s) \times \hat{I}_F(s)) = \delta_1\delta_2 \sum_{s\in G_1(u)}(\mathbf{X}'(s) \times \hat{I}_F(s))$.

To evaluate the model-assisted estimation procedure, we compare four different estimators on the six response variables. For the FOREST variable, we estimate the mean proportion of forest in the region of interest using the following four estimators:

1. EXP, the expansion estimator in (8)
2. PS, a two-phase poststratified estimator with the seven categories of variable NLCD7 as poststrata, representing a standard estimation strategy in FIA
3. REG, a model-assisted estimator as in (10), with parametric regression on the dummy variables for NLCD7 plus linear terms for ELEV90CU, TRASP90, SLP90CU, MRLC00B5, NDVI, Xs, and Ys
4. GAM, the GAM-assisted estimator from (10), with the model described in Section 2.3 fitted by local scoring.

For the remaining five variables, we estimate the means over the region using the first three estimators but replace the fourth (GAM) by the model-assisted estimator that uses the regression model with interactions (15), denoted by REGI.

Table 1 shows the estimates for all six variables (column 3), along with the estimated standard errors (column 4). Following standard FIA practice, these estimated standard errors assume simple random sampling with replacement in phase one and without replacement in phase two. The last column in Table 1 shows the estimated relative efficiency (i.e., the ratio of the estimated variances) of the different estimators compared with GAM or REGI, depending on the variable. These empirical results suggest that the GAM estimator and the related regression estimator with interactions (REGI) dominate the simple expansion estimator, the poststratification estimator, and the paramet-

Table 1. Estimation Results for Northern Utah Mountains Data

Study variable	Estimator	Estimated mean	Estimated standard error	Estimated relative efficiency of GAM/REGI
FOREST (forest/ nonforest binary)	EXP	.51	.02	1.83
	PS	.54	.01	1.38
	REG	.54	.01	1.18
	GAM	.54	.01	
NVOLTOT (total wood volume in ft^3/acre)	EXP	845.81	44.07	1.79
	PS	877.41	39.10	1.41
	REG	877.67	35.35	1.15
	REGI	853.85	32.98	
BA (tree basal area per acre)	EXP	45.19	2.01	1.70
	PS	47.12	1.77	1.33
	REG	47.29	1.63	1.12
	REGI	46.01	1.54	
BIOMASS (total wood biomass in tons/acre)	EXP	13.51	.69	1.96
	PS	14.01	.60	1.51
	REG	14.00	.54	1.19
	REGI	13.60	.49	
CRCOV (percent crown cover)	EXP	21.02	.86	1.73
	PS	22.03	.77	1.39
	REG	22.18	.68	1.09
	REGI	21.64	.65	
QMDALL (quadratic mean diameter in inches)	EXP	3.77	.15	1.26
	PS	3.95	.14	1.08
	REG	3.96	.14	1.01
	REGI	3.89	.14	

NOTE: Estimators are expansion (EXP), poststratification (PS), regression with linear terms for all variables (REG), generalized additive model on the same variables (GAM), regression with linear terms for interactions between GAM predicted forest/nonforest indicator and all variables (REGI). The estimated standard error and relative efficiency (estimated design variance of each estimator divided by estimated design variance of GAM or REGI) use simple random sampling approximation.

ric regression estimator. In particular, the estimated gain in efficiency over the currently used poststratified estimator is $>30\%$ for all variables except QMDALL.

4. VARIANCE ESTIMATION UNDER SYSTEMATIC SAMPLING

4.1 Potential Problems With the Simple Random Sampling Approximation

The estimated efficiencies in Table 1 are somewhat suspect, because they rely on asymptotic variance approximations, and they act as if the actual systematic samples were in fact drawn through simple random sampling. This last point is potentially serious when the number of possible systematic samples is small, as in this 1-in-25 systematic subsample. To illustrate this problem, we reconsider the case of the difference estimator of Section 2.2 when a nonrandom "model" $z^0(s)$ is available for all $s \in G_1(u)$, and assume that the residuals $z(s) - z^0(s)$ are independent normal $(0, \sigma^2)$ random variables. Conditioning on phase one, the model average (over all possible realizations of the normal residuals) systematic sampling variance in (5) is equal to the model average of the simple random sampling variance estimator,

$$\frac{|D|^2}{n_2}\left(1-\frac{1}{h}\right)\left(\sum_{s\in G_2(u,d)}\{z(s)-z^0(s)\}^2 - \left(\sum_{s\in G_2(u,d)}\{z(s)-z^0(s)\}\right)^2 \Big/ n_2\right)\Big/(n_2-1) \quad (16)$$

(see Cochran 1977, thm. 8.5), a fact often used to justify the simple random sampling variance estimator for a population thought to be "random." But such model unbiasedness is not so interesting for a given realization of the population. Indeed, consider

$$F = \frac{\text{systematic sampling variance in (5)}}{\text{simple random sampling variance estimator in (16)}}.$$

Under the foregoing assumptions, it is immediate that this ratio is $\mathcal{F}$-distributed with $h-1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom. As $n_2 \to \infty$,

$$F = 1 + O_p\left(\left\{\frac{2(1+h/n_2)}{h-1}\right\}^{1/2}\right),$$

so that—at least in this simple case—the simple random sampling variance estimator is inconsistent unless h, the number of possible phase-two systematic samples, tends to infinity. In the northern Utah mountains dataset, $n_2 = 968$ but $h = 25$. The quartiles of the corresponding $\mathcal{F}$ distribution are .792 and 1.181. Thus, in about half of the possible realizations of the population, the simple random sampling variance estimator will be off by $\pm 20\%$ or more. The .025 quantile is .514, and the .975 quantile is 1.655, so departures on the order of $\pm 50\%$ are easily possible.

This problem of variance estimation is basically intractable given only the sample, because it amounts to a sample of size 1. Indeed, all of the variance estimators for systematic sampling given by Wolter (1985, sec. 7.2.1) will perform poorly, because they all are forced to rely on within–systematic sample variation to approximate between–systematic sample variation. Therefore, we consider an alternative procedure based on generating a synthetic population.

4.2 Assessing Variability and Efficiency Using a Synthetic Population

Because the standard variance estimators are potentially unreliable in this context, we undertake a numerical experiment to assess the efficiencies of the various estimators. Our approach is to construct a synthetic population that closely mimics the one from which we are sampling, draw all possible systematic samples from that population, and calculate exact design properties of estimators across these samples. This will allow us to assess (1) the performance of the standard variance estimators for a sampling problem with known design variances and (2) the exact design efficiencies of the various estimators for a synthetic population constructed to resemble the real population.

This synthetic population approach represents a departure from generic inference, because it requires specification of a population model for each variable, and the validity of the resulting variance assessment depends on the model's correctness. So, although this procedure would be impractical to implement for all variables in a large-scale survey like the FIA, we believe that performing this type of variance assessment for at least a few key survey variables might be useful to assess the overall reliability of the estimation methodology.

We begin by fitting large parametric models to each variable listed in Table 1. The first model is a logistic regression for the forest/nonforest indicator; it includes six dummy variables for the categories of NLCD7, fourth-order polynomials for

ELEV90CU, TRASP90, SLP90CU, MRLC00B5, NDVI, and the two spatial coordinates, as well as a first-order interaction term for the spatial coordinates. The models for the remaining response variables contain similar terms and are fitted as linear regressions to the positive responses after suitable transformation (typically square root).

Using these fitted models, we create synthetic populations of response variables on all of the phase-one sites. In this procedure, we condition on phase one because its percentage contribution to the empirical variances of the estimators was found to be small, around 5–7% for all six variables, and its contribution is common to all of the estimators; see (7). The simulated FOREST variable is generated with unequal probability Bernoulli random variables using the phase-one covariates, yielding a realistic spatial distribution of forest. The remaining response variables are generated on the transformed scale with Gaussian noise, then mapped back to the original scale. These response variables are set to 0 wherever the simulated FOREST variable is 0.

An alternative procedure for generating the synthetic phase-one sample would be to use the GAM-fitted model that we obtained previously. But we chose not to do this, so as not to bias the results in favor of that estimator. Instead, by generating data from a model for which the GAM only provides an imperfect fit, we expect to be better able to capture the sample-to-sample variability induced by the GAM fitting procedure.

Once the complete phase-one sample is populated as a realization from the foregoing model, we draw all 25 possible phase-two systematic samples, compute the estimated mean and the simple random sampling estimate of the standard deviation for each sample, and then compute averages and variances of these estimates over the 25 samples. Note that these 25 samples represent the entire conditional randomization distribution of the estimators, so that empirical means and variances are exactly the conditional expectation and conditional variance, given phase one. The expectations of the estimators for the synthetic populations (not shown) are comparable to the corresponding estimates for the actual populations given in Table 1. Similarly, the expectations of the simple random sampling standard deviation estimates for the synthetic populations are comparable to the corresponding estimates for the actual populations given in Table 1. These comparisons suggest that the synthetic populations reproduce at least the second-order moment structure of the real data fairly well.

Because these 25 samples constitute all possible samples from the known population, we can evaluate the design bias of the estimators exactly. As expected, the expansion estimator is exactly unbiased for the synthetic population mean, and the remaining estimators are all essentially unbiased (relative biases no more than .25% in all cases) because of their model-assisted structure.

The synthetic population also allows us to evaluate the appropriateness of the simple random sampling approximation for estimating the variance of the estimators obtained under systematic sampling. The last column of Table 2 confirms that these estimated variances are indeed quite unreliable, behaving somewhat like the hypothetical $\mathcal{F}$ random variable described in Section 4.1. Therefore, inference for systematic samples such as those used for the FIA should be done with some caution. Although easy-to-implement but inaccurate generic variance estimators, such as the simple random sampling approximation, are likely to continue to be used in this type of survey, we recommend at least some evaluation based on alternative procedures such as that presented here.

Table 2. Relative Efficiency (design variance of each estimator divided by design variance of GAM or REGI) and Percent Bias of the Simple Random Sampling Variance Estimator

Simulated variable	Estimator	Relative efficiency of GAM/REGI	Percent bias of variance estimator
FOREST (forest/nonforest binary)	EXP	4.51	12.62
	PS	3.13	9.33
	REG	1.92	24.77
	GAM		−31.01
NVOLTOT (total wood volume in ft^3/acre)	EXP	1.31	−23.71
	PS	1.14	−32.11
	REG	1.07	−43.88
	REGI		−52.71
BA (tree basal area per acre)	EXP	2.02	19.35
	PS	1.55	19.81
	REG	1.19	20.97
	REGI		8.93
BIOMASS (total wood biomass in tons/acre)	EXP	1.67	17.48
	PS	1.12	34.31
	REG	1.14	−1.32
	REGI		−14.61
CRCOV (percent crown cover)	EXP	1.49	−4.04
	PS	1.36	−20.93
	REG	1.17	−31.38
	REGI		−44.01
QMDALL (quadratic mean diameter in inches)	EXP	2.55	5.92
	PS	1.79	23.70
	REG	1.20	50.32
	REGI		13.27

NOTE: Values are computed over all 25 possible systematic subsamples from the phase one sample for the synthetic population. Estimators are the same as in Table 1.

Finally, we can exactly evaluate design variances for the five estimators in the synthetic population. Table 2 (column 3) shows the relative efficiencies (design variance of each estimator divided by design variance of GAM or REGI) obtained over the 25 samples. The PS estimator, which is the Forest Service standard, is better than the EXP estimator in all cases, but even the simple regression estimator REG usually offers gains over both the expansion estimator and the PS estimator. The GAM estimator is much more efficient than its competitors for the FOREST variable, and the regression estimator with GAM-dependent interaction terms (REGI) is more efficient than its competitors for all of the other variables. Even though the exact sizes of the efficiency gains differ somewhat, these results confirm those obtained for the different estimators for the original sample shown in Table 1: Real gains are obtained with the GAM-assisted and related regression estimators.

5. CONCLUSION

Auxiliary information from remote sensing or other sources is becoming increasingly available to organizations involved in natural resource surveys. Scientists in these organizations are already developing detailed prediction models for many variables of interest, but they have tended not to use these prediction models in their survey estimation procedures. We have shown

how nonparametric model-assisted estimation techniques can be used to incorporate the results of such modeling efforts in the production of survey estimates. Even in the case of fairly complex models and multiphase designs, estimators can be constructed that are generic, in the sense that they can be easily applied to all variables in a survey and do not depend on a particular model for statistical validity. The overall approach of combining complex models and model-assisted estimation is applicable to a wide range of surveys and can provide large gains in efficiency for relatively little cost.

In the particular application considered in this article, we have provided some theoretical justification for GAM-assisted survey inference in the context of two phases of systematic sampling from a spatial domain. We applied the GAM-assisted methodology in a survey of forest resources in the mountains of northern Utah, a region important for its ecological and land use diversity.

Theoretical properties of this approach in complex surveys, whether using GAM or other nonparametric methods, deserve further investigation. Important open issues include model selection and selection of the smoothing parameters for the nonparametric regression fitting algorithms, because this affects both the estimates of the quantities of interest and their estimated variances.

In the course of this research, the unsatisfactory behavior of the traditional estimator of the design variance under systematic sampling became apparent, and we used an alternative approach, based on a synthetic population, to evaluate our proposed estimation procedure. Future research into these types of alternative variance estimation methods, including choice of models and robustness to their selection, certainly appears to be warranted.

APPENDIX: PROOF OF RESULT 1

Because the estimator $\hat{\theta}$ is unbiased by (4), it suffices to show that its variance goes to 0. By hypothesis, both z and z_0 are Riemann-integrable on D, so that, from (3),

$$\lim_{n_{11},n_{12}\to\infty} \mathrm{E}[\hat{\theta} \mid u] = \theta \quad \text{a.s.},$$

and, from (6),

$$\lim_{n_{21},n_{22}\to\infty} |D| \frac{t_d(u)}{n_{21}n_{22}} = \int_D (z(v) - z^0(v))\, dv \quad \text{a.s.}$$

Because z and z^0 are bounded, we have that

$$\begin{aligned}
&\lim_{n_{11},n_{12}\to\infty} \int_{[0,1]\times[0,1]} (\mathrm{E}[\hat{\theta} \mid u] - \theta)^2\, du \\
&\quad = \int_{[0,1]\times[0,1]} \lim_{n_{11},n_{12}\to\infty} (\mathrm{E}[\hat{\theta} \mid u] - \theta)^2\, du \\
&\quad = 0
\end{aligned}$$

and

$$\lim_{n_{21},n_{22}\to\infty} |D|^2 \frac{\mathrm{E}[S^2(u)]}{(n_{21}n_{22})^2} = |D|^2 \mathrm{E}\left[\lim_{n_{21},n_{22}\to\infty} \frac{S^2(u)}{(n_{21}n_{22})^2}\right] = 0,$$

so that mean squared consistency follows.

[Received March 2003. Revised January 2005.]

REFERENCES

Arnau, X. G., and Cruz-Orive, L. M. (1996), "Consistency in Systematic Sampling," *Advances in Applied Probability*, 28, 982–992.

Bailey, R. G., Avers, P. E., King, T., and McNab, W. H. (eds.) (1994), *Ecoregions and Subregions of the United States* (map), Washington, DC: U.S. Geological Survey.

Breidt, F. J., and Opsomer, J. D. (2000), "Local Polynomial Regression Estimators in Survey Sampling," *The Annals of Statistics*, 28, 1026–1053.

Brewer, K. R. W. (1963), "Ratio Estimation and Finite Populations: Some Results Deducible From the Assumption of an Underlying Stochastic Process," *The Australian Journal of Statistics*, 5, 93–105; corr: 8, 37.

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

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993), "Bias Robust Estimation in Finite Populations Using Nonparametric Calibration," *Journal of the American Statistical Association*, 88, 268–277.

Cochran, W. G. (1942), "Sampling Theory When the Sampling Units Are of Unequal Size," *Journal of the American Statistical Association*, 37, 199–212.

——— (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.

——— (1978), "Laplace's Ratio Estimator," in *Contributions to Survey Sampling and Applied Statistics*, ed. H. A. David, New York: Academic Press, pp. 3–10.

Deville, J.-C., and Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.

Dorfman, A. H. (1992), "Nonparametric Regression for Estimating Totals in Finite Populations," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 622–625.

Dorfman, A. H., and Hall, P. (1993), "Estimators of the Finite Population Distribution Function Using Nonparametric Regression," *The Annals of Statistics*, 21, 1452–1475.

Frayer, W. E., and Furnival, G. M. (1999), "Forest Survey Sampling Designs: A History," *Journal of Forestry*, 97, 4–8.

Frescino, T., Edwards, T. C. Jr., and Moisen, G. (2001), "Modelling Spatially Explicit Structural Attributes Using Generalized Additive Models," *Journal of Vegetation Science*, 12, 15–26.

Fuller, W. A. (2002), "Regression Estimation for Survey Samples," *Survey Methodology*, 28, 5–23.

Gillespie, A. J. R. (1999), "Rationale for a National Annual Forest Inventory Program," *Journal of Forestry*, 97, 16–20.

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

Holt, D., and Smith, T. M. F. (1979), "Post-Stratification," *Journal of the Royal Statistical Society*, Ser. A, 142, 33–46.

Isaki, C., and Fuller, W. (1982), "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89–96.

Jessen, R. J. (1942), "Statistical Investigation of a Sample Survey for Obtaining Farm Facts," Research Bulletin 304, Iowa Agricultural Experiment Station.

Kim, J.-Y. (2004), "Nonparametric Regression Estimation in Survey Sampling," unpublished doctoral thesis, Iowa State University.

Kuo, L. (1988), "Classical and Prediction Approaches to Estimating Distribution Functions From Survey Data," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 280–285.

Moisen, G., and Edwards, T. (1999), "Use of Generalized Linear Models and Digital Data in a Forest Inventory of Utah," *Journal of Agricultural, Biological and Environmental Statistics*, 4, 372–390.

Moisen, G. G., and Frescino, T. S. (2002), "Comparing Five Modelling Techniques for Predicting Forest Characteristics," *Ecological Modelling*, 157, 209–225.

Opsomer, J. D. (2002), "Nonparametric Regression Model," in *Encyclopedia of Environmetrics*, Vol. 3, eds. A. H. El-Shaarawi and W. W. Piegorsch, Chichester, U.K.: Wiley, pp. 421–427.

Robinson, P. M., and Särndal, C.-E. (1983), "Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling," *Sankhyā*, Ser. B, 45, 240–248.

Royall, R. M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," *Biometrika*, 57, 377–387.

Särndal, C.-E. (1980), "On π-Inverse Weighting versus Best Linear Unbiased Weighting in Probability Sampling," *Biometrika*, 67, 639–650.

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

Scott, C. T., Bechtold, W. A., Reams, G. A., Smith, W. D., Hansen, M. H., and Moisen, G. G. (2004), "Sample-Based Estimators Utilized by the Forest Inventory and Analysis National Information Management System," in *The Enhanced Forest Inventory and Analysis Programnational Sampling Design and Estimation Procedures*, eds. W. A. Bechtold and P. L. Patterson, Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station, pp. 43–68.

U.S. Department of Agriculture Forest Service (1992), "Forest Service Resource Inventories: An Overview," technical report, Washington, DC.

Vogelmann, J. E., Howard, S. M., Yang, L., Larson, C. R., Wylie, B. K., and Driel, N. V. (2001), "Completion of the 1990s National Land Cover Data Set for the Conterminous United States From Landsat Thematic Mapper Data and Ancillary Data Sources," *Photogrammetric Engineering and Remote Sensing*, 67, 650–662.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

Wright, R. L. (1983), "Finite Population Sampling With Multivariate Auxiliary Information," *Journal of the American Statistical Association*, 78, 879–884.

Wu, C., and Sitter, R. R. (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information From Survey Data," *Journal of the American Statistical Association*, 96, 185–193.

Comment

David RUPPERT

This stimulating article is an ingenious combination of survey methods and "mainstream statistics." The authors use nonparametric regression to improve estimates from natural resource surveys while maintaining the traditional inferential methods of survey methodology than rely on randomization of the sample rather than on a model. This is quite an achievement, and I congratulate them. There is certainly too little contact between survey sampling and the rest of statistics, probably because many statisticians lack knowledge of sampling theory. Thus an article such as this combining the two areas is a welcome addition to the literature.

The authors' results clearly demonstrate that the model-assisted estimator $\widehat{\theta}_{\text{ma}}$ works well in this example. In a case study such as this, the main goal is, of course, to find a methodology suitable for the problem at hand, but it is natural to wonder whether the methodology can be recommended as a general-purpose tool. This discussion addresses two related questions: "When does $\widehat{\theta}_{\text{ma}}$ work well relative to $\widehat{\theta}_{\text{exp}}$ (and vice versa)?" and "why does the model-assisted estimator work very well in this case study?"

The authors' result 1 assumes that $z(s)$ and $z_0(x)$ are continuous almost everywhere. Translating mathematical assumptions such as this into something operational is challenging. Should we think of FOREST as continuous almost everywhere or not? This question may not be fully answerable, but I believe that for practical purposes, FOREST is discontinuous, and the patchiness of forests is the main reason why the GAM/REGI methodology that the authors develop is successful in their case study. Assuming that z is continuous, the expansion estimator can be written as

$$\widehat{\theta}_{\text{exp}} = \sum_{s\in G_2(u,d)} \frac{z(s)}{1/(\delta_1\delta_2 h)} = \sum_{s\in G_1(u)} \frac{z\{w(s)\}}{1/(\delta_1\delta_2)}$$
$$\approx \sum_{s\in G_1(u)} \frac{z(s)}{1/(\delta_1\delta_2)} \approx \theta, \tag{1}$$

where $w(s)$ is the point on the coarse grid in the same cell as s. Therefore, the accuracy of $\widehat{\theta}_{\text{exp}}$ should be greatest when z is very smooth, so that $z(s)$ is close to $z\{w(s)\}$ and $\sum_{s\in G_1(u)} \frac{z(s)}{1/(\delta_1\delta_2)}$ is close to θ.

The model-assisted estimator can be written as

$$\widehat{\theta}_{\text{ma}} = \sum_{d'} \sum_{s\in G_2(u,d')} \left\{ \frac{\widehat{\mu}\{\mathbf{X}(s)\}}{1/(\delta_1\delta_2)} + \frac{z(s) - \widehat{\mu}\{\mathbf{X}(s)\}}{1/(\delta_1\delta_2)} \frac{I_{\{d=d'\}}}{1/h} \right\}$$
$$= \sum_{s\in G_1(u)} \left\{ \frac{\widehat{\mu}\{\mathbf{X}(s)\}}{1/(\delta_1\delta_2)} + \frac{z\{w(s)\} - \widehat{\mu}[\mathbf{X}\{w(s)\}]}{1/(\delta_1\delta_2)} \right\}.$$

This estimator is accurate when the model achieves good predictions, so that $\widehat{\mu}\{\mathbf{X}(s)\}$ is close to $z(s)$ and $\widehat{\mu}[\mathbf{X}\{w(s)\}]$ is close to $z\{w(s)\}$. Smoothness of z is not required for accuraty.

To compare the expansion and model-assisted estimators, we simulated a one-dimensional example for simplicity. The simulation model was $z(s) = \sin\{k\pi s(1 + 2s^2)\}$, where k is either 2 (smooth) or 5 (less smooth). Here "smoothness" refers not to the number of derivatives (which is infinity) but rather to how fast the function oscillates. When sampled on a fixed grid, an oscillatory function will appear discontinuous if the oscillations are sufficiently rapid compared with the spacings between grid points. Moreover, $x(s) = \text{logistic}\{k_2(z(s) - .5)\} + \sigma\epsilon$, $\epsilon \sim \text{N}(0, 1)$, where k_2 is fixed at 2 and σ is either .01 (model rather accurate) or .1 (model less accurate). The regression of $z(s)$ on $x(s)$ was estimated by a penalized spline (Ruppert, Wand, and Carroll 2003). The coarse grid had 20 points; the fine grid, 100 points.

We start with the case where $k = 2$ and $\sigma = .01$. The performances of the expansion and model-assisted estimators for one simulated dataset are shown in Figure 1. Boxplots of the absolute errors of the expansion and model-assisted estimators for 100 simulated datasets are shown in Figure 2. In this case, the expansion estimator is better than the model-assisted estimator.

Next, we look at the case where $k = 5$ and $\sigma = .01$ in Figures 3 and 4. Because k is large, z oscillates rapidly and the expansion estimator is less accurate than when $k = 2$. Because here, as in the previous case, σ is small (.01), the model-assisted estimator is rather accurate and, as shown in Figure 4, more accurate than the expansion estimator.

Finally, we look at the case where $k = 5$ and $\sigma = .1$ in Figures 5 and 6. Because k is again large, z the expansion esti-

David Ruppert is Professor of Engineering, School of Operations Research & Information Engineering, Cornell University, Ithaca, NY 14853 (E-mail: *dr24@cornell.edu*). This research was supported by National Institutes of Health grant CA57030 at Texas A&M University.

Journal of the American Statistical Association
June 2007, Vol. 102, No. 478, Applications and Case Studies
DOI 10.1198/016214506000001509

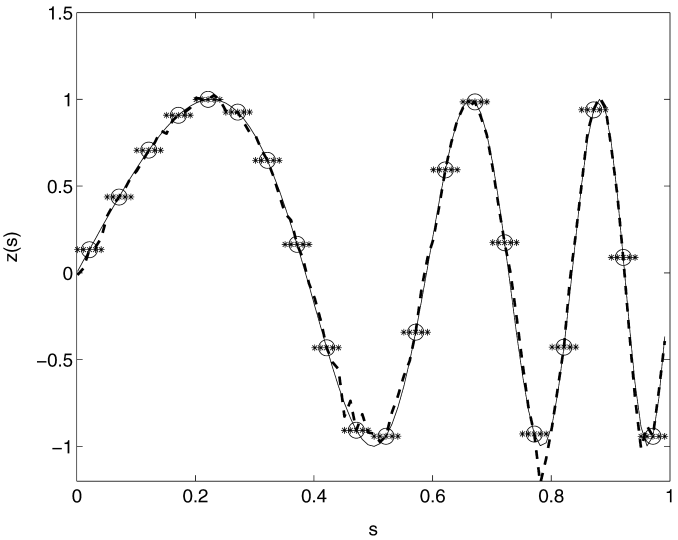

Figure 1. Results for One Dataset for the Case Where $k = 2$ and $\sigma = .01$. The solid line is the process z. The large circles show the evaluation of z on the coarse grid. The asterisks show the values on the fine grid of the estimate of z used in the expansion estimator; that is, $z(s)$ is estimated by $z\{w(s)\}$, and the expansion estimator is proportion to the sum of $z\{w(s)\}$ over the fine grid. The dashed line is the model-based estimate of z. The dashed line is close to the solid line, demonstrating that the model-based predictions are rather accurate in this case. The model-assisted estimator is proportional to the sum of these predictions over the fine grid.

mator is less accurate than when $k = 2$. But because σ is now large, the model-assisted estimator is somewhat inaccurate and, as shown in Figure 6, less accurate than the expansion estimator.

In summary, we can expect the model-assisted estimator to perform well relative to the expansion estimator when z is discontinuous, or at least highly oscillatory, and z can be predicted accurately from x. It seems likely that both of these conditions are true in the authors' case study.

The REGI estimator is an interesting adaptation of a general methodology (model-assisted estimator) to particular features

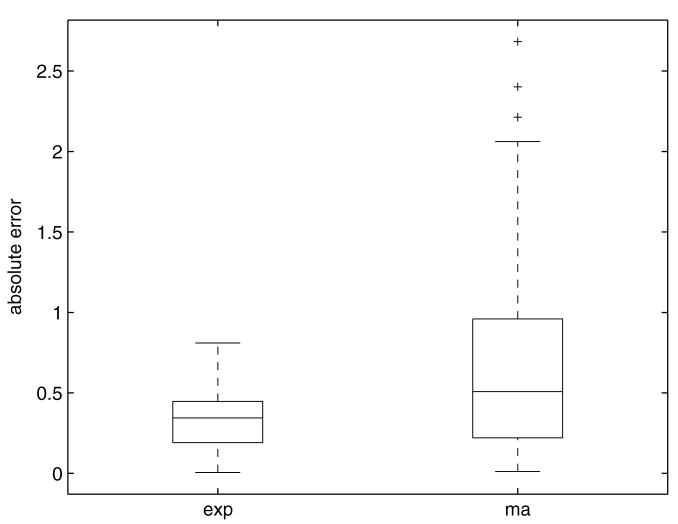

Figure 2. Boxplots of the Absolute Errors on the Expansion and Model-Assisted Estimators Showing That in the Case Where $k = 2$ and $\sigma = .01$, the Expansion Estimator Is More Accurate Than the Model-Based Estimator.

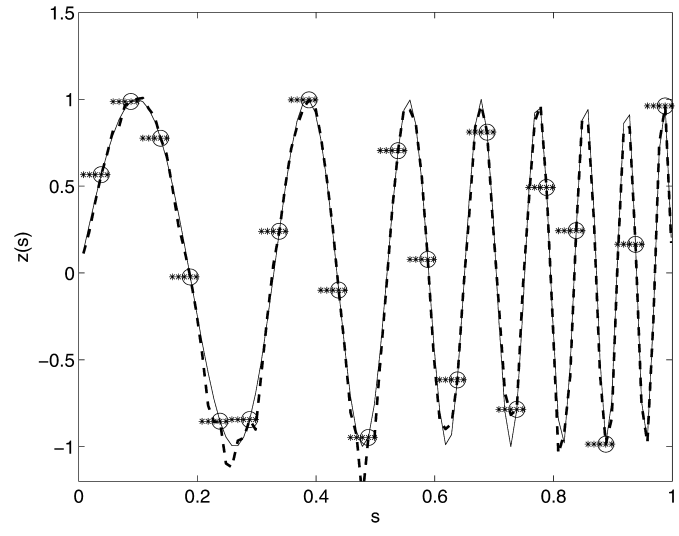

Figure 3. Results for One Dataset for the Case Where $k = 5$ and $\sigma = .01$. The solid line, large circles, asterisks, and dashed line are the same as in Figure 1. The dashed line is close to the solid line, demonstrating that the model-based predictions are rather accurate in this case.

of this case study (FOREST is discontinuous). In essence, it uses one model to predict the presence or absence of forest and a second model to predict variables of interest, such as total wood volume, *but only where forest is predicted to exist*.

I close by suggesting another estimator. This starts with the same model as used by REGI to predict the presence/absence of forest. Then, for those s where forest is predicted to exist, it estimates $z(s)$ by $z\{w(s)\}$, where z is, for example, total wood volume. Because this methodology combines model-assisted estimation with the expansion estimator, it could be called the "hybrid estimator." I would be very interested to see how well this hybrid estimator performs on the authors' case study. It should perform well when FOREST is discontinuous, but variables such as total wood volume are continuous *when restricted to forested areas*.

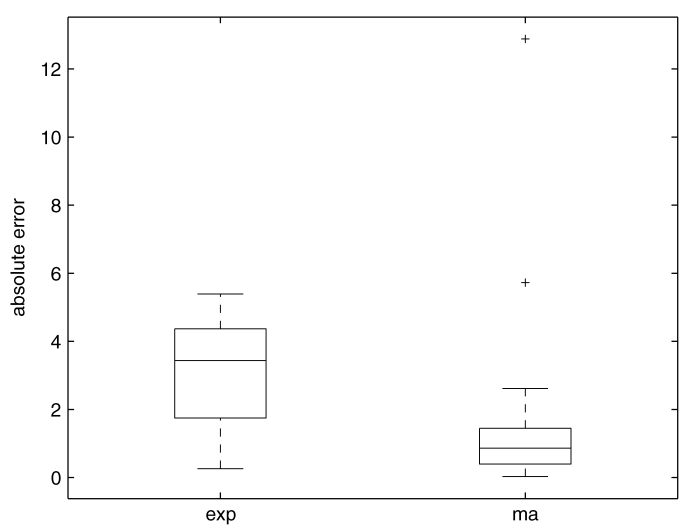

Figure 4. Boxplots of the Absolute Errors on the Expansion and Model-Assisted Estimators Showing That in the Case Where $k = 5$ and $\sigma = .01$, the Model-Assisted Estimator Is More Accurate Than the Expansion Estimator.

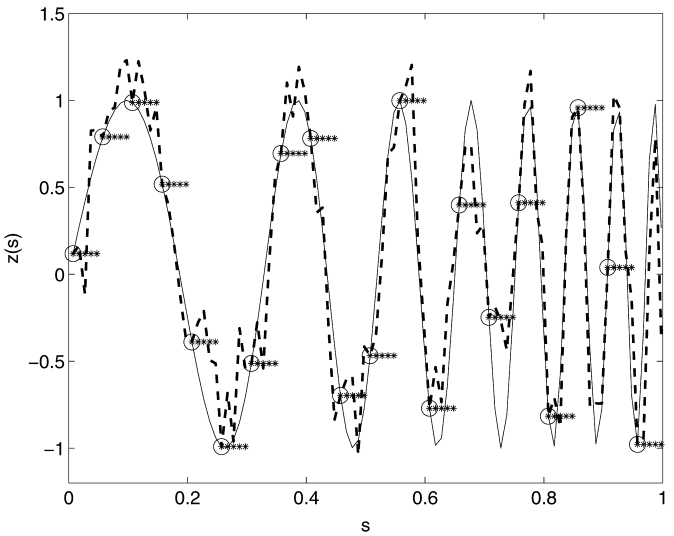

Figure 5. Results for One Dataset for the Case Where $k = 5$ and $\sigma = .1$. The solid line, large circles, asterisks, and dashed line are the same as in Figure 1. The dashed line is far from the solid line, demonstrating that the model-based predictions are not very accurate in this case.

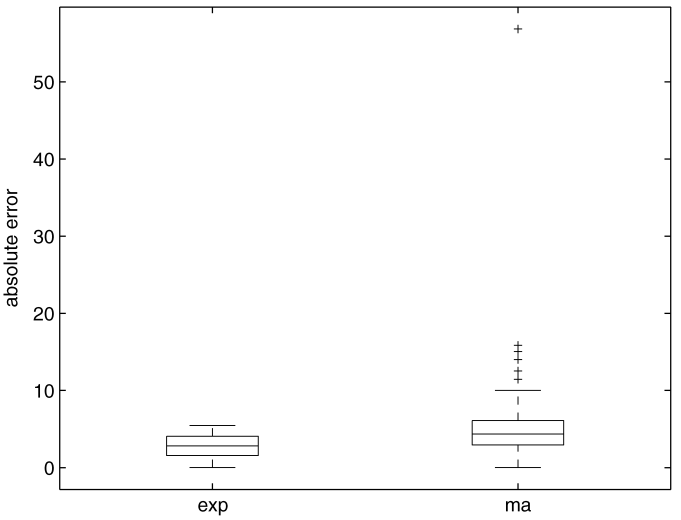

Figure 6. Boxplots of the Absolute Errors on the Expansion and Model-Assisted Estimators Showing That in the Case Where $k = 5$ and $\sigma = .1$, the Expansion Estimator Is More Accurate Than the Model-Assisted Estimator.

Clearly, there is no uniformly most accurate estimator; thus we need methods for choosing among the expansion, model-assisted, and hybrid estimators when working with a particular case study.

ADDITIONAL REFERENCE

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

Comment

Mary C. CHRISTMAN

Opsomer et al. have provided an interesting example extending the two-phase model-assisted generalized difference estimator (Sarndal et al. 1992) to the case in which the regression component is replaced by a generalized additive model (GAM) estimator. Then the result of that model is used for estimating other response variables. One of the response variables (FOREST) is modeled using a GAM, and the GAM estimator of presence/absence of FOREST then is used as an explanatory variable in the regression component of Sarndal et al.'s model-assisted generalized difference estimator for five other response variables. This approach follows the model calibration method of Wu and Sitter (2001), in which the GAM estimator used as a predictor for other variables is treated as fixed with respect to the design. Such a method allows construction of survey weights applicable for all response variables of interest. The authors argue that the resulting estimators for the survey variables are approximately linear, at least asymptotically.

To assess the efficiency of this approach to model-assisted estimation, the authors compare their model estimators to three other estimators, one based on only phase-two data (ESP), one based on poststratification (PS), and one based on Sarndal et al.'s model-assisted estimator using linear regression (REG). The main distinguishing feature between the authors' GAM estimator and the REG estimator is that they add an additional predictor variable based on the GAM modeling of FOREST that modifies the regression model used to estimate the response variables. They argue—and rightly so—that absence of FOREST forces the other response variables to equal 0 and that an appropriate model should recognize that fact. Because they are interested primarily in whether GAM modeling provides better efficiency than a regression approach, it would have been more informative had they compared their GAM estimator with a regression-type estimator that was otherwise identical to the GAM estimator except with FOREST estimated using a logistic regression model in place of the GAM.

From table 2, it appears that predicting presence/absence of FOREST is more efficient when using a GAM than when using

Mary C. Christman is Associate Professor, Department of Statistics, University of Florida, Gainesville, FL 32611 (E-mail: *mcxman@ufl.edu*).

Journal of the American Statistical Association
June 2007, Vol. 102, No. 478, Applications and Case Studies
DOI 10.1198/016214506000001518

a logistic regression, but whether this translates into more efficient prediction of the other survey variables is unclear, because the REG estimators for those other variables do not incorporate predictions of the presence or absence of FOREST. The GAM approach likely still would be more efficient than the analogous REG estimator with FOREST predictions incorporated, but this would depend on the form of the logistic regression used to predict FOREST. In fact, the authors used a logistic regression model later in their approach to estimating variance that might have been an appropriate model for comparing the REG and GAM/REGI estimators.

Unfortunately, assessing the variance of systematic samples is difficult when only one systematic sample is taken. In fact, table 1 reports the variance as though the systematic samples were taken as simple random samples in both phases. The authors recognize that this is inappropriate and take advantage of the two-phase nature of their sampling design to develop a technique that they call the "synthetic" approach to estimating the systematic variance for comparing the different estimators. The authors estimate the variance of the phase-two systematic sample conditioned on the phase-one sample. They argue that the contribution of the phase-one sampling is small relative to the sampling variance of phase two, and thus this is an appropriate alternative to ignoring the systematic nature of the sampling and treating the data as arising from random sampling. Now, ignoring the sampling variability of the phase-one sample is equivalent to treating the predicted value at each phase-one sampling location as the block average for the grid cell (1 km $\times$ 1 km) centered on the (x, y) coordinates of the location. If it is true that the contribution of the phase-one sampling to the total variance is in fact small, then this is a reasonable approach. If not, and the scale of spatial variability is finer than the grid distances in the phase-one sample, then the estimators are underestimating the true systematic sampling variance. One way to determine whether the variance is being underestimated would be to take advantage of the data used in the survey. The predictor variables were obtained from digital elevation models (DEM) and 30-m resolution thematic mapper (TM) imagery and thus can be determined for locations other than the sampled phase-one systematic sample locations. Thus it would be possible to apply the authors' approach to creating a synthetic population over different phase-one samples to determine the contribution of the phase-one sample to the overall variance.

Assuming the authors' argument that the contribution of phase one to the empirical variances of the estimators is small and constant across all estimators, the synthetic approach would be valid for comparing the different estimators if the variance of the GAM estimator was being estimated appropriately. The variance of the GAM estimator is underestimated, because it fails to incorporate the variance of the predictor of FOREST in the estimators of the five survey variables in which it is used. Hence, although the GAM estimator is likely more efficient than the other approaches, it has not yet been shown to what degree this is true. The authors have made a good first pass at the comparison through their synthetic approach, but there remains much to be done to determine better estimators of the variance for two-phase systematic sampling designs in general and of the GAM estimators used here in particular.

The GAM always outperforms the other estimation approaches, as expected when additional informative auxiliary data are used, but what is fascinating is the bias of SRS estimator of systematic variance for the different response variables (table 2). The SRS (design-based) estimator of design-based systematic variance is positively or negatively biased as the intraclass correlation

$$\rho = \sum_{i=1}^{p} \sum_{j=1}^{n} \sum_{j' \neq j} \frac{(z_{ij} - \mu)(z_{ij'} - \mu)}{pn(n-1)\sigma^2}$$

is less than or greater than $-1/(N-1)$ where N is population size, n is the systematic sample size, p is the number of possible systematic samples of size n, μ is the population mean, and z_{ij} is the value of the jth unit in the ith systematic sample. In the example given by Opsomer et al., table 2 clearly shows that the intraclass correlation varies significantly across both the response variables and the estimators. This implies that the effect of model-assisted estimation depends on which variable is under consideration and, more importantly, that the choice of model assistance induces an intraclass correlation that then affects the efficiency of the estimator through this correlation.

Overall, the authors have provided an interesting and detailed example of how GAMs might be incorporated into model-assisted estimation and have provided impetus for further study into the possible use of systems of simultaneous equations for estimating a suite of response variables such as those of interest to the Forest Service.

Comment

Roderick J. LITTLE

1. INTRODUCTION

I appreciate the editor's kind invitation to discuss this interesting article and also the choice of an environmental sampling topic for the Applications and Case Studies invited paper. Sampling methods are a key contribution of statistics to science, and these days are not as well represented in *JASA* as I would like.

Roderick J. Little is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *rlittle@umich.edu*).

Journal of the American Statistical Association
June 2007, Vol. 102, No. 478, Applications and Case Studies
DOI 10.1198/016214506000001527

Moreover, environmental monitoring is currently a hot topic in more senses than one, and the article concerns a real-world problem with fascinating statistical issues.

I congratulate Opsomer, Breidt, Moisen, and Kauermann (henceforth OBMK) on their progress in improving the forest inventory survey estimates. There seems to be good evidence that their methods result in gains in precision over previous alternatives. The article continues these authors' useful previous contributions to robust survey modeling.

My discussion moves from the general to the particular. In Section 2 I consider viewpoints on survey inference, comparing the authors' philosophy with my own. In Section 3 I discuss the systematic sampling design of the Forest Inventory Survey, challenges in variance estimation for this design, and possible alternatives. In Section 4 I provide some general comments on OBMK's model-assisted estimation procedure. Finally in Section 5 I make some more specific comments on the choice of regression models in this setting.

2. MODES OF SURVEY INFERENCE

The inferential approach adopted in the article might be described as "quasi-design-based, model-assisted" inference. The approach is model-assisted in that OBMK use models to improve the efficiency of the inferences, while basing the inference on design-based properties in repeated sampling. Design-based, model-assisted inference is perhaps the prevailing mode of survey inference these days, and is guided by the reasonable aim of attempting to capture strengths of model and design approaches, as popularized by the work of Särndal and colleagues (e.g., Särndal et al. 1992). The "quasi" part (which I admit makes me somewhat "queasy") arises because, as OBMK clearly explain, design-based inference is not possible for the systematic sample design, because the design variance cannot be estimated. This leads to assessing variance estimates under other designs and for a population simulated under a particular model. Because the standard errors for the latter are model-based, we have model-based standard errors for design-based model-assisted inference, a combination that I must say leaves my head spinning!

These approaches reflect what to me is a form of "inferential schizophrenia" shared by many survey samplers, who tend to be design-based for some problems, like inference for overall population quantities from large probability samplers, and model-based for other problems, like nonresponse or small-area estimation. This is all very pragmatic but is unprincipled, because design-based and model-based inference have conflicting features, and to me it is illogical to simultaneously subscribe to both theories.

I prefer a unified approach to survey inference that can be applied to all problems (Little 2004). This approach might be termed "Bayesian model-based, design-assisted." The inference is model-based and Bayesian but design-assisted in that key features of the design are incorporated in the model to convey robustness and stop it from going astray because of gross specification errors. Bayesian model-based inference for finite population quantities Q is based on the posterior predictive distribution of Q for a Bayesian model, and essentially involves predicting values of nonsampled elements and propagating uncertainty in those predictions. Models need to be robust to design features such as sampling weights, stratification and clustering, and inferences should be "calibrated" in the sense of having good repeated-sampling properties, such as design consistency (Little 2006).

With large samples, I like models that make weak assumptions about functional forms, as in models involving splines (Breidt and Opsomer 2000; Zheng and Little 2003, 2004, 2005). Priors should be multilevel to incorporate clustering and limit subjective features. In many cases, this approach yields inferences similar to "superpopulation models" but with simpler interpretations.

3. DESIGN–BASED INFERENCE FOR THE FOREST INVENTORY SURVEY

As OBMK note, design-based inference is not possible for systematic samples. Here randomization rests solely on two sets of draws, $U = (u_1, u_2)$, $0 < u_j < 1$, and $D = (d_1, d_2)$, $1 \le d_j \le 5$, which shift the phase-one grid and phase-two subgrid in north/south and east/west directions. Design-based SEs cannot be computed, because there is only one realization of U and D—replication is needed to compute a randomization variance. OBMK consider two fixes for this problem that assume a different design and hope that the resulting answers are not seriously wrong. They assume simple random sampling at both phases, but comparisons with the inferences on the simulated population appear to indicate that this assumption is not very reasonable. More stratified alternative designs yield better approximations to systematic designs, as discussed by Wolter (1985); for example, one might create strata based on four adjacent sampled sites and compute the variance using replication methods that drop one site and weight up the other three. OMBK's approach based on simulating an artificial population is ingenious, but obviously model-based and hence inferentially schizophrenic. If one is really serious about design-based inferences, then it seems preferable to modify the design to make it possible by introducing replications of U and D, rather than pretending that the inference is design-based when it really is not.

More generally, I am skeptical about the value of randomization for this systematic design based on just two random draws. Different choices of U and D shift *all* of the sites in one direction—so $U = (.1, .1)$, $D = (1, 1)$ versus $U = (.9, .9)$, $D = (5, 5)$ might have a systematic effect on overall estimates if latitude and/or longitude affect the outcome of interest. In repeated samples, an "extreme" (U, D) is balanced by other choices of U in other hypothetical samples—this is the promise of design consistency. But why are other possible samples that are not chosen relevant? Why not avoid an extreme choice and center the grid and subgrid with $U = (.5, .5)$, $D = (3, 3)$? Although this no longer is a probability sample, it surely is a better way to limit latitude/longitude effects. The legitimacy conveyed by the random selection of a single (U, D) seems to me phony—a sacrifice to the gods of randomization! A design with considerably more than two random selections is needed to realize the desired balancing properties of randomization.

Meaningful design-based inferences involve confidence intervals, not standard errors; an estimated standard error is only as good as the coverage of its derived confidence interval. Suppose that the phase-one grid is fixed and that quantities of interest are defined to be the population average values on that grid. Second-stage sampling yields only 25 possible repeated samples. Thus for any procedure, only a finite number of coverages are possible—$0/25, 1/25, \ldots, 25/25$—that is, a 95% confidence interval does not exist! The frequentist interpretation of the usual interval—an estimate plus or minus two standard errors—becomes increasingly problematic as the number of possible samples is reduced. These problems do not arise in credibility intervals under the Bayesian paradigm.

A final comment on the design is that here all sites have same selection probability. This is at the other extreme from what environmental scientists like to do: sample "interesting areas" purposively. A statistically principled compromise is to stratify sites and oversample interesting ones with higher probability. One criterion for "interesting" might be local heterogeneity in the variables of interest. I wonder whether that approach is worth considering in the forestry surveys considered here.

4. MODEL–ASSISTED ESTIMATION VERSUS ROBUST MODEL–BASED ESTIMATION

The predictions from the regression models adopted by OBMK are "calibrated" by adding the sample mean of the residuals (design-weighted if there are design weights). The rationale is to ensure design consistency (Särndal et al. 1992). I prefer adopting regression models that yield design consistent estimates without the calibration step. This is very easy to do, as discussed by Firth and Bennett (1998). Examples of such models include all linear models that include the intercept term and all generalized linear models with canonical links that include the intercept.

Therefore, OBMK's models can be made to yield design-consistent estimates without calibration by simply adding an intercept. One might think this is the same as calibration, but this is not the case. Consider the simplest model with a single X and no intercept, $y_i \sim_{\text{ind}} \mathrm{N}(\beta x_i, \sigma^2)$. Calibration yields predictions of the form $\bar{y} + \hat{\beta}(x_i - \bar{x})$, which have the same form as predictions from the model with an intercept $y_i \sim_{\text{ind}} \mathrm{N}(\alpha + \beta x_i, \sigma^2)$. However, the calibration estimator fits a nonzero intercept, but then estimates the slope assuming that the intercept is zero. This seems to me a very strange thing to do! When an intercept is being fitted, surely it is more sensible to estimate the other regression coefficients with the intercept included in the model, thus preserving the optimality of the predictions under the assumed model. Does the calibration approach have any known advantage over modifying the model in the manner suggested? (For other examples in which calibration yields strange estimates, see Little 1983.)

5. SPECIFIC REGRESSION MODELING COMMENTS AND SUGGESTIONS

In justifying the use of GAMs, OBMK state that

> "survey" (i.e., regression) weights... can be used for any variables collected in the same survey, and to the extent that they follow (the assumed model), they will benefit from the efficiency gain. Therefore, it is desirable to specify the model as flexibly as possible.

This statement conveys the impression that there is nothing to lose by modeling flexibly, but including predictors that are not predictive adds variance, as in the case of poststratification (Holt and Smith 1979). Thus throwing the "kitchen sink" into the model has a downside. Furthermore, the choice of GAMs sacrifices interactions for flexible main effects—for example, it seems doubtful that the shape of the response surface over the spatial coordinates is the same for all vegetation classes, but this is assumed in OBMK's models. Zheng and Little (2003, 2004, 2005) avoid the so-called "curse of dimensionality" of nonparametric regression by confining the nonparametric part to the relationship with the sampling weight. Because this is an equal probability design, this tactic would lead to parametric models (with intercept) in this setting, which, however, retain the design consistency property.

OBMK note that the weights from regression are same for all outcomes, if the same set of covariates is included in all models. In practice, one may want to tailor models for particular outcomes (or sets of related outcomes). I believe that the importance of having the same set of regression weights is overstated, given modern computing power.

The OBMK application involves outcomes that are meaningful in forested locations but effectively zero in locations that are not forested. They deal with these outcomes by interacting the covariates with a variable $\hat{I}_F$ indicating whether the predicted probability of FOREST is above an estimated overall probability from the expansion estimator, which is about .5. Thus $\hat{I}_F = 0$ means that the probability of being forested is below average, not zero. A common two-stage modeling strategy in such settings that seems preferable is a logistic regression on the indicator for whether or not a sample site is forested, followed by linear regression for the forest measures, restricted to sample sites observed to be forested. Different prediction models may well be useful for these two steps.

Finally, the bivariate smooth fit on spatial coordinates captures large-scale spatial structure, but local spatial structure also may be important. Specifically, do residuals from the model show any evidence of spatial correlation, and does including X's from phase-one sites neighboring the sample sites improve fit? A simple approach to assessing the latter is to derive a best predictor $\hat{X}$ by the methods discussed by OBMK, and then include in the model values of $\hat{X}$ for phase-one sites neighboring the sampled site as well as for the sampled site itself.

ADDITIONAL REFERENCES

Firth, D., and Bennett, K. E. (1998), "Robust Models in Probability Sampling," *Journal of the Royal Statistical Society*, Ser. B, 60, 3–21.

Little, R. J. A. (1983), "Estimating a Finite Population Mean From Unequal Probability Samples," *Journal of the American Statistical Association*, 78, 596–604.

——— (2004), "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling," *Journal of the American Statistical Association*, 99, 546–556.

——— (2006), "Calibrated Bayes: A Bayes/Frequentist Roadmap," *The American Statistician*, 60, 213–223.

Zheng, H., and Little, R. J. (2003), "Penalized Spline Model-Based Estimation of the Finite Population Total From Probability-Proportional-to-Size Samples," *Journal of Official Statistics*, 19, 99–117.

——— (2004), "Penalized Spline Nonparametric Mixed Models for Inference About a Finite Population Mean From Two-Stage Samples," *Survey Methodology*, 30, 209–218.

——— (2005), "Inference for the Population Total From Probability-Proportional-to-Size Samples Based on Predictions From a Penalized Spline Nonparametric Model," *Journal of Official Statistics*, 21, 1–20.

Rejoinder

Jean D. OPSOMER, F. Jay BREIDT, Gretchen G. MOISEN, and Göran KAUERMANN

We thank the editor for organizing the discussion of our article and the discussants for their interesting and insightful contributions.

1. RESPONSE TO RUPPERT

Ruppert created an interesting simulation experiment to compare the efficiency of the expansion estimator and the nonparametric model-assisted estimator. His results clearly show that the results depend on the relative goodness of fit of the model estimator $\hat{\mu}\{\mathbf{X}(s)\}$ and the "expansion estimator" $z\{w(s)\}$, as estimators of the target function $z(s)$. If $z\{w(s)\}$ is a good approximation to $z(s)$ (the situation in figs. 1 and 2), then model-assisted estimation is unnecessary and can degrade the fit, especially in situations involving challenging model fits. A qualitatively similar situation is illustrated in figures 5 and 6, where the increased noise in the model further degrades the goodness of fit of the model relative to the naive approximation $z\{w(s)\}$. In the third situation considered by Ruppert, the model-assisted estimator improves the precision of the survey estimator when the model can be estimated accurately, whereas $z\{w(s)\}$ is not a good approximation for $z(s)$. This case is illustrated in figures 3 and 4.

Note that in the first two cases, the average efficiency of the expansion and the model-assisted estimators are still close, with the latter negatively affected by a small number of dramatic "model failures" as shown in the boxplots. It would be up to the statistician computing model-assisted survey weights to guard against this type of problem, by performing appropriate model-fitting and weight diagnostics. Assuming that these extreme cases can be successfully avoided, then the overall conclusion of Ruppert's experiment (and indeed, much of survey practice) is that model-assisted estimation has efficiency that is either close to that of direct (expansion) estimation or potentially much better, if the right model is found. Nevertheless, these results clearly indicate that model selection is critical in survey estimation. To date, formal methods for this purpose are mostly lacking in survey statistics and certainly represent an interesting avenue for future research.

Ruppert's idea of combining the expansion estimator and the predicted presence/absence of forest in a "hybrid estimator" would provide a good alternative to the REGI procedure that we applied, if most of the predictive power of the model comes from its ability to separate forested areas from unforested areas. As described, Rupert's estimator is a *model-based* predictor, because it has the form

$$\hat{\theta}_{\text{hyb,mb}} = \sum_{s\in G_1(u)} \frac{z\{w(s)\}\hat{I}_F(s)}{1/(\delta_1\delta_2)}.$$

We computed this estimator on the suite of variables NVOLTOT, BA, BIOMASS, CRCOV, and QMDALL and found the estimates to be systematically lower—sometimes much lower—than the other estimate displayed in table 2. The reason for this can be seen by considering an alternative, *model-assisted* version of Ruppert's hybrid, which combines the foregoing model-based hybrid with a design bias adjustment,

$$\begin{aligned}\hat{\theta}_{\text{hyb,ma}} &= \sum_{s\in G_1(u)} \frac{z\{w(s)\}\hat{I}_F(s)}{1/(\delta_1\delta_2)} + \sum_{s\in G_2(u,d)} \frac{z(s) - z(s)\hat{I}_F(s)}{1/(\delta_1\delta_2 h)} \\ &= \hat{\theta}_{\text{exp}} + \sum_{s\in G_1(u)} \frac{z\{w(s)\}\hat{I}_F(s)}{1/(\delta_1\delta_2)}\left(1 - \frac{I_{s\in G_2(u,d)}}{1/h}\right).\end{aligned}$$

By arguments given in our article, this model-assisted hybrid should be approximately design-unbiased. The bias adjustment contains summands that are 0 if the site is not forested, so $z(s) = 0$, or if the site is forested and the model correctly predicts that the site is forested. Otherwise, if the model *incorrectly* predicts that the site is not forested, then the summand is $z(s) > 0$. Thus a *positive* bias adjustment is needed for the model-based hybrid estimator. With this bias adjustment, the model-assisted hybrid estimator performs similarly to REGI, with differences due to the fact that REGI works harder at predicting $z(s)$ on the phase-one grid.

2. RESPONSE TO CHRISTMAN

Christman argues that a parametrically specified logistic regression model for the forest/nonforest indicator would provide an appropriate comparison for the GAM we used. A carefully chosen parametric model (constructed from, say, low-degree polynomials) certainly would be able to capture the trends observed in figure 2 and is in fact likely to result in a model-assisted estimator that is more efficient than the nonparametric

Jean D. Opsomer is Professor, Department of Statistics, Iowa State University, Ames, IA 50011 (E-mail: *jopsomer@iastate.edu*). F. Jay Breidt is Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523. Gretchen G. Moisen is Research Forester, U.S. Department of Agriculture Forest Service, Rocky Mountain Research Station, Ogden, UT 84401. Göran Kauermann is Professor, Department of Economics, University of Bielefeld, 33501 Bielefeld, Germany.

In the Public Domain
Journal of the American Statistical Association
June 2007, Vol. 102, No. 478, Applications and Case Studies
DOI 10.1198/016214506000001536

version. However, one of the important advantages of the nonparametric approach is its flexibility, which implies that it is able to capture patterns like those in figure 2 without having to prespecify a parametric form. The price paid for this flexibility in the model-assisted context is a (typically modest) reduction in efficiency (see the simulation experiments in Breidt and Opsomer 2000).

As noted by Christman, the synthetic population approach will provide a valid measure of estimator uncertainty only if the assumed model for the population is (approximately) correct. Although we took care to model the relationship between the phase-one and phase-two variables, we assumed that the simple random sampling approximation for the phase-one variance was satisfactory, because it was estimated to represent only a small portion of the overall variance and was unaffected by the various regression estimation approaches. But a more complete treatment of this problem could indeed be undertaken, and the availability of "wall-to-wall" auxiliary information would make modeling of the phase-one variance contribution possible.

For a given finite population and set of systematic samples, the intraclass correlation provides a convenient measure for the bias of the simple random sampling variance approximation. Ignoring the effect of model fitting, the systematic sampling variance and hence the intraclass correlation depend on the behavior of the $z(s) - z^0(s)$ [see (5) and (6)], which depend on both the variable z itself and the model assumed in z^0, as noted by Christman. An interesting observation based on the synthetic population results in table 2 is that whereas modeling, whether based on poststratification, linear regression, or nonparametric regression, has a significant effect on the bias of the variance approximations, it does not seem to result in an overall bias reduction.

3. RESPONSE TO LITTLE

Today's survey statisticians, both methodologists and practitioners, must deal with increasingly complex data available from different sources and of different reliability, and use these data not only to produce summary tabular information for populations and domains (descriptive inference), but also to fit complicated statistical models (analytical inference) to address various hypotheses. Little describes many of the current approaches combining model-based and design-based inference as "inferential schizophrenia," "pragmatic but unprincipled." But these approaches do have clear guiding principles, which can be summarized as follows:

- *Make descriptive inferences as model free as they can be.* For large probability samples, no model is needed; for small area estimation, models are essential for borrowing strength across related domains. Much of survey practice lies between these two extremes, and it is here that nonparametric modeling can be particularly useful.
- *Make analytical inferences that properly reflect complex design features.* The "messy" nature of the approaches used reflects both the messy nature of the data-generating process and the degree to which the various sources of uncertainty involved (e.g., sampling design, nonresponse, measurement error, population model) can be reliably described.

In the end, whether a resulting approach is mostly design-based, mostly model-based, or a true hybrid, the key feature of valid inference based on survey data is that the characteristics of the design must be accounted for. A range of approaches is being developed by researchers (including Little and ourselves), and this is likely to remain an area of active research for many years to come.

Systematic sampling is often used in surveys of spatial domains, where it has some attractive characteristics, including simplicity and intuitiveness, equal probability, and approximate self-balancing over broad domains of interest. This design also suffers from a number of serious drawbacks, however, including the lack of a design-based variance estimator, as well as several other effects that are consequences of the low degrees of freedom, as noted by Little. Many solutions have been proposed for variance estimation and for extending the sampling design to circumvent this estimation problem (see Wolter 1985, chap. 7, for a review). One possible, albeit somewhat "schizophrenic" (in Little's terminology), approach, described by Wolter (1985), consists of estimating the *anticipated variance* (i.e., the model-expected design variance) under a specific model for the population. When this model is correct, it is generally possible to estimate the anticipated variance, and this estimate then can be used as a measure of the uncertainty of the design-based estimator. This is conceptually similar to the synthetic population approach that we used in our forestry application. Clearly, both approaches are only as good as the model used to represent the population and require careful attention to model selection and model fit.

We agree with Little that using models that result in self-calibration (i.e., models that produce calibrated predictions without having to add the sample mean of the residuals) is desirable, and in fact for many regression models, the model-assisted estimator is exactly equal to the design-weighted model prediction (see Särndal et al. 1992, result 6.5.1). Because our final model uses the covariate vector in (15), which does not contain an intercept, it is not self-calibrated. However, we refitted the final model with an intercept, and found that each of the resulting estimates changed by $<2\%$.

As noted by Little and clearly illustrated by Ruppert, a poor choice of model can lead to model-assisted estimators that are actually less efficient than expansion estimators. The choice of the GAM in our application was motivated by the desire to use a parsimonious multivariate model while maintaining the ability to represent the complex relationship between the forest/nonforest indicator and the covariates. In the second step, the choice of a parametric model for the other variables involving interactions with the predicted forest indicators was similarly influenced by a desire to balance parsimony and comprehensiveness. More generally, the current paradigm of generic inference for survey estimators, which dictates a single set of weights to be interpreted as "adjusted sampling weights" for all survey variables, makes model selection a particularly challenging problem.