MODEL-ASSISTED ESTIMATION OF FOREST RESOURCES WITH GENERALIZED ADDITIVE MODELS

J.D. Opsomer, G.G. Moisen, J.Y. Kim
J.D. Opsomer, Iowa State University, Ames, IA 50011, U.S.A.

Key words: multi-phase survey estimation, nonparametric regression, local scoring, calibration.

Abstract:

Multi-phase surveys are often conducted in forestry, with the goal of estimating tree characteristics and volume over large regions. 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 exact relationship between the ground visit measurements and the remote sensing variables is not known, hence it is modelled using generalized additive models. Nonparametric estimators for these models are discussed and applied to forest data collected in northeastern Utah in the United States. By using these model predictions in a model-assisted survey estimation procedure for tree volume and related variables, we improve the accuracy of the survey estimates compared to currently available estimation procedures. The procedures described in this article are applicable to many other survey contexts.

1 Introduction

Accurate estimation of forest resources over large geographic areas is of significant interest to forest managers and forestry scientists. In nationwide forest surveys of the U.S., design-based estimates of quantities like total tree volume, growth and mortality, or area by forest type are produced on a regular basis. In the current article, we consider the estimation of such quantities within a 3.18 million ha ecoregion in northeastern Utah. Figure 1 displays the region of interest and the sample points collected in the early 1990's for the survey we will consider here. While this article will focus on this particular example, the approach proposed here can be applied in other forest and natural resource estimation problems as well.

Currently, forest survey estimates are being produced through a two-phase sampling procedure, with phase one consisting of aerial photo-based information collected on an intensive sample grid, and phase two consisting of a subset of that grid visited in the field. Photo-interpreted vegetation cover type and ownership are often used for stratification (or post-stratification) of phase two field points, and design-based estimates of population totals are then calculated. Because of the increasing availability of a wide variety of inexpensive information derived from satellite observations, there is a tremendous opportunity for effective methods to merge forest inventory data with diverse auxiliary information, to both reduce costs and further improve precision on forest survey estimates.

At the same time, scientists within government agencies and in other institutions are working on developing predictive and analytical models describing forest characteristics. Because of the multivariate nature of the data and the incomplete understanding of the relationships between variables, nonparametric and semiparametric models are often found to be a good compromise between model specification and flexibility. Such modelling efforts would in principle provide a good source of auxiliary information to produce more precise survey estimators. The purpose of this article is to propose an approach for including such models into the design-based estimation framework, and to apply it to the Utah forest dataset.

Model-assisted survey estimation (Särndal *et al.* [7]) is a well-known approach for incorporating auxiliary information in design-based survey estimation. It assumes the existence of a "superpopulation model" between the auxiliary variables and the variable of interest for the population to be sampled. The estimation of population quantities of interest is then performed in such a way that the design properties of the estimators can be established. This is in contrast to purely model-based estimation, for which no design-based inference is possible.

While model-assisted estimation has the potential to improve the precision of survey estimators when appropriate auxiliary information is available, it typically requires that these models be linear or at least have a known parametric shape. Breidt and Opsomer [1] introduced local polynomial regression

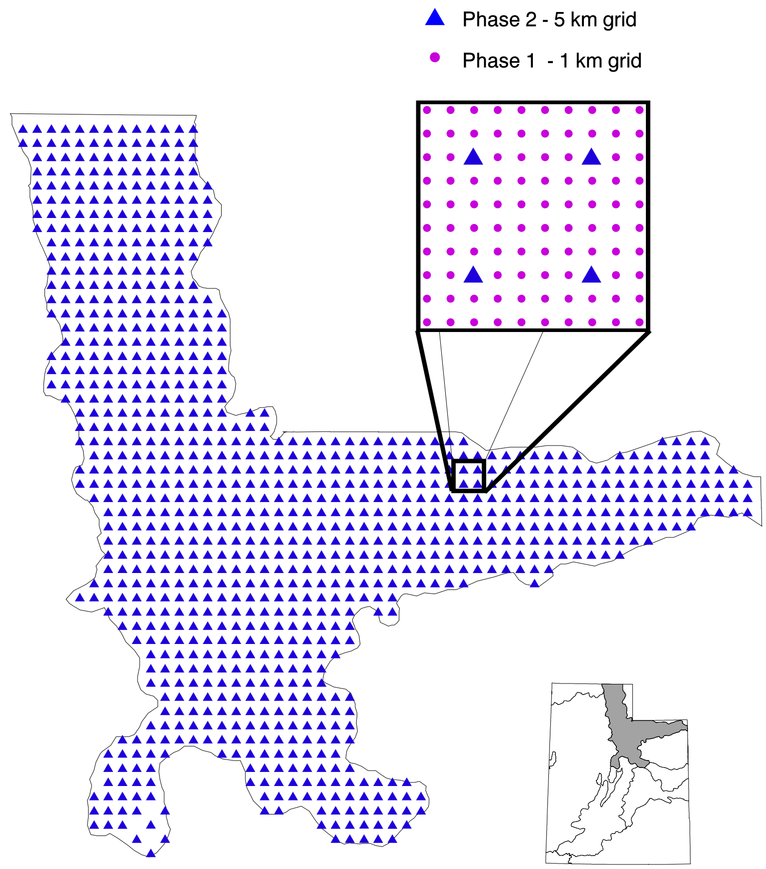

Figure 1: Representation of study region in northeastern Utah. Each dot represents a field-visited sample plot. See Section 4 for an explanation of the phase 1 and phase 2 plots.

estimation, a survey estimation approach combining the modelling flexibility of nonparametric regression with model-assisted estimation. In this article, we describe how this approach can be used with survey data from a multi-phase design and with generalized nonparametric regression models.

In Section 2, we review model-assisted survey estimation for multi-phase designs, and in Section 3, we apply it to the generalized additive model. Section 4 discusses the models used for prediction for the Utah forest inventory. In Section 5, we show the results of applying nonparametric model-assisted estimation methods to the Utah data.

2 Model-assisted Nonparametric Estimation

The Utah data was collected in two phases on a regularly spaced grid. Such data is often analyzed as two phases of simple random sampling without replacement. This section describes the estimators considered for this application.

Let $U = \{1, \ldots, i, \ldots, N\}$ represent the population of elements to be sampled, $s_a \subseteq U$ the sample of elements selected in phase 1 and $s \subseteq s_a$ the sample of elements selected in phase 2. In this application, the elements are plots of approximately 1 acre (0.405 ha) in size. Let n_a and n represent the phase 1 and phase 2 sample sizes, respectively. Finally, let y_i represent a generic variable of interest for element i.

The design-based estimator for the population total $t_y = \sum_U y_i$ is

$$\hat{t}_y = \frac{N}{n} \sum_s y_i, \tag{1}$$

The variance of this estimator is composed of two components, one for each phase of sampling, and can be written as

$$\begin{aligned}\text{Var}(\hat{t}_y) &= N^2 \left(1 - \frac{n_a}{N}\right) \frac{S^2_{yU}}{n_a} \\ &\quad + N^2 \text{E}_a \left(\left(1 - \frac{n}{n_a}\right) \frac{S^2_{ys_a}}{n} \right),\end{aligned}$$

where S^2_{yU} and $S^2_{ys_a}$ are the variances of the y_i in the population and in the phase 1 sample, respectively, and E_a is the expectation with respect to the first phase of sampling. An unbiased estimator for the variance is

$$\hat{V}(\hat{t}_y) = N^2\left(1-\frac{n_a}{N}\right)\frac{S^2_{ys}}{n_a} + N^2\left(1-\frac{n}{n_a}\right)\frac{S^2_{ys}}{n}, \quad (2)$$

with S^2_{ys} the variance of the y_i in s. More details on two-phase sampling and the estimator are given in Särndal *et al.* [7], chapter 9.

Suppose now that we would like to use auxiliary information to improve on the precision of the above estimator. We assume that the auxiliary information is available for all the elements in the phase 1 sample s_a, but not for the remaining elements in the population (this is the most common situation in two-phase sampling). Let $\boldsymbol{X}_i$ represent a vector of auxiliary variables that is observed for all $i \in s_a$. The superpopulation model

$$\mathrm{E}(Y_i|\boldsymbol{X}_i) \equiv \mu_i = f(\boldsymbol{X}_i) \quad (3)$$

is assumed for the population. Traditionally, $f(\cdot)$ is assumed to be linear. Recently, Wu and Sitter [8] studied more general parametric regression models, and Breidt and Opsomer [1] considered nonparametric models.

The model is fitted on the data $(\boldsymbol{X}_i, y_i)$, $i \in s$ using a nonparametric regression technique. A wide range of such method are available, including kernel-based methods, spline methods and orthogonal decomposition-based approaches (see e.g. Opsomer [6] for an overview), and in principle any of these can be used to produce predictions $\hat{\mu}_i$ for all $i \in s_a$. In model-assisted estimation, the estimator $\hat{t}_y$ in (1) is then replaced by

$$\hat{t}_{ym} = \frac{N}{n_a}\left(\sum_{s_a}\hat{\mu}_i + \frac{n_a}{n}\sum_{s}(y_i - \hat{\mu}_i)\right). \quad (4)$$

The design-based estimator $\hat{t}_y$ is replaced by model predictions over the phase 1 sample, corrected by the observed deviations between the true values and model predictions over the phase 2 sample, appropriately weighted up to the whole population.

When $\hat{\mu}_i$ is obtained from an appropriately weighted linear regression, $\hat{t}_{ym}$ is known as the *generalized regression estimator* (Särndal *et al.* [7]). In that case, $\hat{t}_{ym}$ is design consistent, regardless of whether the model (3) is correctly specified or not, and its design variance can be approximated by

$$\mathrm{Var}(\hat{t}_{ym}) \approx N^2\left(1-\frac{n_a}{N}\right)\frac{S^2_{yU}}{n_a} + N^2\mathrm{E}_a\left(\left(1-\frac{n}{n_a}\right)\frac{S^2_{es_a}}{n}\right), (5)$$

where $S^2_{es_a}$ is the variance of the residuals $e_i = y_i - \hat{\mu}_i$ for the model fit on $(y_i, \boldsymbol{X}_i)$ for $i \in s_a$. Similarly, a variance estimator is found by replacing S^2_{ys} by S^2_{es} in (2).

The poststratified estimator for t_y can be considered as a special case of the generalized regression estimator, in which the auxiliary variables are categorical. By classifying the phase 2 sample into a small number of post-strata based on the phase 1 information, this estimator is often used as a relatively simple way to incorporate auxiliary information in the estimation. See Särndal *et al.* [7] for explicit expressions for the post-stratified estimator and its variance.

It is intuitively clear why a model can improve the efficiency of the estimator. If the model fits the data well, the variance of the residuals e_i can be expected to be smaller than the variance of the y_i. If the model fits poorly, the residual variance should be equally large or even potentially larger than the measurement variance. Hence, the efficiency gains of the model-assisted estimator depend on the selection of a good model for $f(\cdot)$ in (3). The same reasoning applies when $f(\cdot)$ is fitted by nonparametric regression (Breidt and Opsomer [1]). In that case, the estimator $\hat{t}_{ym}$ in (4) can be more efficienct even even when the specific form of $f(\cdot)$ is unknown.

The estimator $\hat{t}_{ym}$ has some additional desirable properties. If the nonparametric regression method is a *linear smoothing method*, in the sense that $\hat{\mu}_i = \sum_s s_{ij}y_j$ for a set of smoothing weights s_{ij} that do not depend on the y_j, $\hat{t}_{ym}$ can be written as a linear combination of the sample observations $\hat{t}_{ym} = \sum_s w_i y_i$, with weights w_i independent of the y_i. These regression weights can be used for any variables collected in the same survey, and to the extent that such variables also follow model (3), they will also benefit from the efficiency gain. When (local) polynomial regression is used for model fitting, the resulting regression weights are *calibrated* for the auxiliary variables in phase 1, so that $\sum_s w_i\boldsymbol{X}_i = \frac{N}{n_a}\sum_{s_a}\boldsymbol{X}_i$.

3 Model-assisted Estimation Using Generalized Additive Models

Suppose now that $f(\boldsymbol{X}_i)$ is the generalized additive model

$$\mathrm{E}(Y_i|\boldsymbol{X}_i) \equiv \mu_i = g(m_1(\boldsymbol{X}_{1i})+\ldots+m_D(\boldsymbol{X}_{Di})) \quad (6)$$

for some known link function $g(\cdot)$ and unknown smooth functions $m_d(\cdot)$, $d = 1, \ldots, D$, where the $\boldsymbol{X}_{di}$ are known subsets of the vector $\boldsymbol{X}_i$. Given a set of estimated functions $\hat{m}_d(\cdot), d = 1, \ldots, D$, model predictions $\hat{\mu}_i = g^{-1}(\hat{m}_1(\boldsymbol{X}_{1i}) + \ldots + \hat{m}_D(\boldsymbol{X}_{Di}))$ and corresponding residuals $e_i = y_i - \hat{\mu}_i$ are readily calculated, for instance using the `gam()` local scoring estimation routines (Hastie and Tibshirani [3]) implemented in *S-Plus*. Assuming consistency of the generalized additive model estimators, the estimator $\hat{t}_{ym}$ will be design consistent under conditions similar to those in Breidt and Opsomer [1]. The variance approximation (5) is then again appropriate for this case.

Unless the link function $g(\cdot)$ is the identity link, local scoring estimators are not produced by linear smoothers. Hence, the resulting estimator $\hat{t}_{ym}$ is no longer a linear combination of the y_i. In addition, the inverse transformation $g^{-1}(\cdot)$ used to calculate the $\hat{\mu}_i$ introduces bias in the model portion of the estimator. Both of these disadvantages can be corrected, however, by introducing an additional regression estimation step. To do this, we perform a linear regression of the y_i on the $\hat{\mu}_i$. Let $\hat{\mu}_i^*$ represent the fitted values from this new regression. The estimator

$$\hat{t}^*_{ym} = \frac{N}{n_a}\left(\sum_{s_a}\hat{\mu}_i{}^* + \frac{n_a}{n}\sum_{s}(y_i - \hat{\mu}_i^*)\right) \quad (7)$$

is corrected (up to first order) for the bias introduced by the link function, and it can again be written as a linear combination of the observations y_i for $i \in s$. Note that this new estimator is now calibrated for the model fits $\hat{\mu}_i$ in phase 1, not for the auxiliary variables themselves. Wu and Sitter [8] proposed a calibration adjustment for their nonlinear regression estimators that is similar to this second regression step.

4 Generalized Additive Models for the Forest Inventory Data

In this section, we discuss fitting generalized additive models to the Utah forest inventory. Data used in this study were collected on a 5 km sample grid as illustrated in Figure 1. On each of 968 phase 2 sample plots, numerous forest site variables and individual tree measurements were collected, including a binary classification of the plot as "forest" or "nonforest" (FOR), and a measure of total wood volume (NVOL) expressed in cuft per acre. In addition to the field plot data, remotely sensed auxiliary information was extracted on the 5 km field plot locations as well as on an intensified 1 km grid (25,386 points), which will represent the phase 1 data. These data came from four sources: (1) elevation (ELEV), a transformed aspect (TRASP), and slope (SLOPE) from digital elevation models produced by the U.S. Defense Mapping Agency; (2) a Normalized Difference Vegetation Index (NDVI) from a biweekly Advanced Very High Resolution Radiometer composite; (3) vegetation cover type from the U.S. National Land Cover Data (NLCD1) based on a 30-m resolution Thematic Mapper imagery collapsed to seven vegetation classes; and (4) spatial coordinates (X and Y).

Moisen and Edwards [5], Moisen [4] and Frescino *et al.* [2] developed parametric and nonparametric models relating remotely sensed data to forest attributes observed during field visits. Taking a similar approach here, response variables FOR and NVOL were modeled as nonparametric functions of ancillary predictor variables X, Y, ELEV, TRASP, SLOPE, NDVI, and NLCD1 through generalized additive models. Logit and log link functions $g(\cdot)$ were chosen for the FOR and NVOL models, respectively. The model (6) was fitted 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 [6] for an explanation of the loess smoothing method and smoothing parameter selection). Predictor variables ELEV and NDVI entered the model as univariate smooth terms, while X with Y, and TRASP with SLOPE contributed as bivariate smooth functions. NLCD1 entered as a categorical variable in both models. The plots of the smooth contributing terms in the FOR model are shown in Figure 2. The plots for NVOL are not shown but are qualitatively similar.

5 Design-based Survey Estimation for the Forest Inventory Data

We calculate the following estimators for FOR and NVOL:

1. the design-based estimator $\hat{t}_y$ in (1)

2. the model-assisted estimator $\hat{t}_{ym}$ in (4)

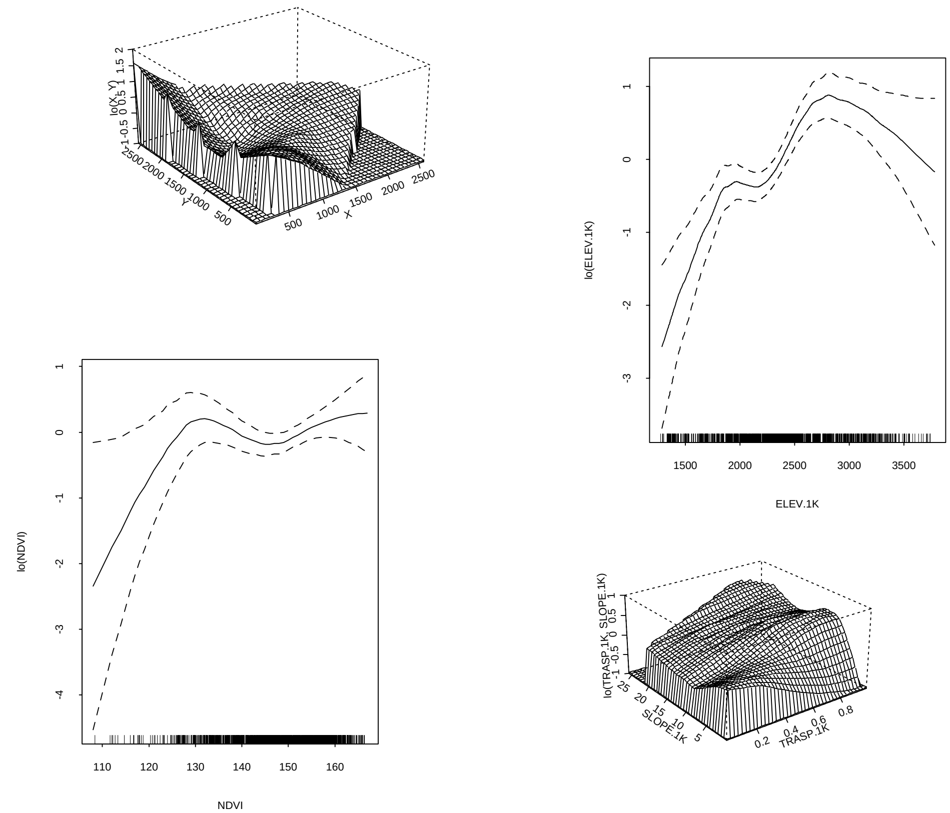

Figure 2: GAM model fits for FOR variable

3. the bias-corrected model-assisted estimator $\hat{t}^*_{ym}$ in (7)

4. a two-phase poststratified estimator with the seven categories of variable NLCD1 as post-strata, denoted by $\hat{t}_{yps}$.

Table 1 shows the estimates for both FOR and NVOL, as well as the estimated standard deviations, when the model is fit using local scoring.

Estimator	FOR (in 000)	NVOL (in 000,000)
1. $\hat{t}_y$	3,389 (99)	5,587 (271)
2. $\hat{t}_{ym}$	3,389 (81)	5,540 (224)
3. $\hat{t}^*_{ym}$	3,390 (80)	5,540 (224)
4. $\hat{t}_{yps}$	3,375 (86)	5,614 (245)

Table 1: Estimates for FOR and NVOL (estimated standard deviations in parentheses).

There is a 17–18% decrease in the estimated standard deviations for the model-assisted estimator $\hat{t}_{ym}$ for both variables, compared to the purely design-based estimator $\hat{t}_y$. The bias-correction step does not appear to give any additional precision benefit, but is still worthwhile because it provides weights.

The traditional method for improving the precision of design-based estimators relies on post-stratification, which in this case provides an improvement of 13% for FOR and 10% for NVOL. Since the post-stratification variable, NLCD1, was included as a categorical variable in the generalized additive models, $\hat{t}_{ym}$ can be interpreted as a further refinement of $\hat{t}_{yps}$. Modelling appears to provide a small but valuable increase in precision of 4% for FOR and 8% for NVOL.

Survey weights, obtained directly from the design as inverse inclusion probabilities or after adjustment from a regression, are useful primarily because they can be applied to any variable of interest in a survey. To illustrate this point, we can calculate regression weights by fitting the variable FOR on the $\hat{\mu}_i$ obtained from the generalized additive model for that variable, and use these weights to estimate the total for NVOL. Conversely, NVOL-derived regression weights can be obtained and used to estimate the total of FOR. The first two lines of Table 2 show

Estimator	FOR (in 000)	NVOL (000,000)
FOR weights	3,390 (80)	5,611 (239)
NVOL weights	3,365 (87)	5,540 (224)
FOR/NVOL weights	3,390 (80)	5,544 (224)

Table 2: Estimates for FOR and NVOL using different regression weights (estimated standard deviations in parentheses).

the estimates obtained using both regression weights and their estimated standard deviations.

Clearly, weights specifically calculated for one variable do not achieve the full precision improvement when applied to the other variable. A simple solution to recapture the full efficiency improvement for the main variables in a survey is to perform the regression step simulatneously using all the predictions for these variables as covariates. In this case, this means performing a multiple regression on both the FOR and the NVOL predictions, so that the regression weights are jointly calibrated to both model fits. Line three in Table 2 shows the estimates produced after this multiple regression step. The results for both variables are now almost identical to those found by calibrating them individually to their respective predictions, achieving the desired results.

6 Conclusion

Auxiliary information from remote sensing or other sources is becoming increasingly available to organizations involved in natural resource surveys. In this article, we have explained how nonparametric model-assisted estimation techniques can be used to incorporate such auxiliary information in the production of survey estimates, even in the case of fairly complex models and multi-phase designs. This was illustrated using generalized additive models in a survey of forest resources in Northeastern Utah.

The theoretical properties of this approach in complex surveys are currently being investigated by the authors. An important open issue concerns the selection of the smoothing parameters for the nonparametric regression fitting algorithms, since this directly affects both the estimates of the quantity of interests and their estimated variances.

References

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

[2] T.S. Frescino, T.C. Edwards, Jr, and G.G. Moisen. Modelling spatially explicit structural attributes using generalized additive models. *Journal of Vegetation Science*, 12:15–26, 2001.

[3] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, Washington, D. C., 1990.

[4] G.G. Moisen. *Comparing nonlinear and nonparametric modelling techniques for stratification and mapping in forest inventories of the Interior Western USA*. PhD thesis, Utah State University, 2000.

[5] G.G. Moisen and T.C. Edwards, Jr. Use of generalized linear models and digital data in a forest inventory of utah. *Journal of Agricultural, Biological and Environmental Statistics*, 4:372–390, 1999.

[6] J. D. Opsomer. Nonparametric regression in environmental statistics. In *Encyclopedia of Environmetrics*. Wiley, 2001. To appear.

[7] C.-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.

[8] C. Wu and R. R. Sitter. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185–193, 2001.