PROPERTIES OF THE ENDOGENOUS POST-STRATIFIED ESTIMATOR USING A RANDOM FORESTS MODEL

John Tipton, Jean Opsomer, and Gretchen G. Moisen[1]

Abstract.—Post-stratification is used in survey statistics as a method to improve variance estimates. In traditional post-stratification methods, the variable on which the data is being stratified must be known at the population level. In many cases this is not possible, but it is possible to use a model to predict values using covariates, and then stratify on these predicted values. This method is called endogenous post-stratification estimation (EPSE). In this paper, we investigate methods to automatically select the number of post-strata for EPSE. We do this in the context of models fitted by Random Forests with the stratum boundaries set at quantiles of the predicted distribution.

INTRODUCTION

Post-stratification is used in survey sampling designs as a method to improve variance estimates by calibrating to known population quantities (Särndal et al. 2003). In the U.S. Forest Service Forest Inventory Analysis Program (FIA), the stratum categories are often land cover classifications based on remote sensing data. In many cases it is desirable to use the FIA data itself to develop the very maps used for stratification. This method is called endogenous post-stratification estimation (EPSE) and the theoretical properties of this method were first introduced by Breidt and Opsomer 2008, and then extended to a broader class of nonparametric models by Dahlke et al. 2012. The use of FIA data to construct maps and then the subsequent use of these maps as post-stratum to construct estimates has the potential for substantially reducing variance in these estimates. Given the increased use of more complex predictive models in developing forest attribute maps in survey applications, there is an urgent need for simulation studies to investigate the properties of the EPSE method and determine under which conditions the EPSE estimator works and under which it fails.

This paper has three main goals. First is to compare the EPSE estimator properties using a linear model, a spline model, and a Random Forests model (Breimam 2001) to develop post-stratum maps. The second goal is to investigate the effects of using estimated stratum boundaries instead of fixed stratum boundaries. The third aim of the simulation study is to investigate the effects of a minimization of the variance estimate on the EPSE estimator.

THE ENDOGENOUS POST-STRATIFIED ESTIMATOR

Following the EPSE framework described by Briedt and Opsomer (2008), a sample s of size n is taken from a population U = $\{1, \dots, i, \dots, N\}$ of size N according to a probability design $p(\cdot)$ where $p(s)$ is the probability of drawing the sample s. For each $i \in U$ a vector of covariates x_i and a response y_i is observed. There is assumed to be a true relationship between x_i and y_i, denoted $m(\cdot)$, where $E[y_i \mid x_i] = m(x_i)$ which is estimated by $\hat{m}(x_i)$.

[1] Graduate Student (JT), 102 Statistics Building, Colorado State University, Fort Collins, CO 80523; Statistics Department Chair (JO), Colorado State University; Research Forester (GGM), U.S. Forest Service, Rocky Mountain Research Station. JT is corresponding author: to contact, call 970-231-9494 or email at Jtipton25@gmail.com.

The models used for this simulation study are a linear model, a spline model, and a Random Forests model. Details on the models and dataset are omitted from this paper. In EPSE, the model predictions $\hat{m}(x_i)$, $i = 1, \ldots, N$ are sorted into H fixed stratum based on the stratum boundaries $\tau_1, \tau_2, \ldots, \tau_{H-1}$ where $\hat{m}(x_i)$ is in the h^{th} stratum if $I(\tau_{h-1} < \hat{m}(x_i) \leq \tau_h) = 1$. The estimated sample counts in stratum h are given by $\hat{n}_h$ and the estimated population counts in stratum h are given by $\hat{N}_h$. The ratio $\frac{\hat{n}_h}{\hat{N}_h}$ is the estimated stratum weight for stratum h. Using the estimated stratum weights, the stratum mean $\hat{\mu}_h$ is calculated for each stratum h by

$$\hat{\mu}_h = \frac{1}{\hat{n}_h} \sum_{i \in s} y_i I_{\{\tau_{h-1} < \hat{m}(x_i) \leq \tau_h\}}. \qquad [1]$$

The EPSE estimator $\hat{\mu}_y$ for the population mean is calculated by

$$\hat{\mu}_y = \frac{1}{N} \sum_{h=1}^{H} \hat{N}_h \hat{\mu}_h. \qquad [2]$$

The estimates $\hat{V}(\hat{\mu}_y)$ for $Var(\hat{\mu}_y)$ are calculated using the post-stratified formulas in Särndal et al. 2003 by

$$\hat{V}(\hat{\mu}_y) = (1 - \frac{n}{N}) \frac{1}{N^2} \sum_{h=1}^{H} \hat{N}_h^2 \frac{s_h^2}{\hat{n}_h} \qquad [3]$$

where $s_h^2 = \frac{1}{\hat{n}_h - 1} \sum_{i \in s} [(y_i - \hat{\mu}_h)^2 I_{\{\tau_{h-1} < \hat{m}(x_i) \leq \tau_h\}}]$ is the sample variance for stratum h. For the simulation study, fixed stratum values for τ_h are considered following Breidt and Opsomer (2008). Also, estimated stratum values $\hat{\tau}_h$ based on quantiles of the model predictions for the set of population covariates x_i, $i \in U$ are considered.

EMPIRICAL PROPERTIES OF THE ESTIMATOR

The data used for this study are from the pilot study of Utah for the 2011 National Land Cover Data (NLCD) canopy cover map. The study region consisted of 4151 observations in Utah. At each location, aerial photography was interpreted to determine the percentage canopy cover. This is the forest response variable of interest in these simulations. Empirical models of tree canopy cover were then derived by modeling this tree canopy cover as functions of Landsat TM reflectance values and topographic values described in detail in Coulston et al. (2012). These empirical models were then used to predict tree canopy cover and develop the post-strata through a variety of binning rules applied to predicted values.

Comparison of Fixed vs. Estimated Stratum Boundaries

The first simulation study is designed to address two questions. First, how does the EPSE performance compare between the linear model, spline model, and Random Forests model when the stratum boundaries are fixed. This is the case where the theory is well known. And second, how is the EPSE performance affected when the stratum boundaries are estimated by sample quantiles as compared to using predetermined fixed stratum boundaries. To get the fixed stratum boundaries, each empirical model was fit using the full dataset. Assuming the empirical model is correct, the population quantiles of percentage canopy cover are determined. This stratification scheme is fixed and not dependent on the sample (i.e. it is the same value for all samples, or a priori). The strata for the fixed stratum boundaries simulation are $(-\infty, Q_1(y)]$, $(Q_1(y), Q_2(y)]$, $(Q_2(y), Q_3(y)]$, $(Q_3(y), \infty)$ where $Q_1(y)$ represents the 25th percentile of the predicted values of percentage tree cover based on the full population. The second stratification uses estimated quantiles from the empirical model fit using the sample and then the estimated quantiles are used as stratum boundaries. Note that in this second scheme, the quantiles are dependent on the sample s, or a posteriori. The strata for the estimated stratum boundaries simulation are $(-\infty, \hat{Q}_1(y)]$, $(\hat{Q}_1(y), \hat{Q}_2(y)]$, $(\hat{Q}_2(y), \hat{Q}_3(y)]$, $(\hat{Q}_3(y), \infty)$ where $\hat{Q}_1(y)$ represents the 25th percentile of the predicted values of percentage tree cover based on the full population. For both of these strata definitions the behaviour of the EPSE estimator $\hat{\mu}_y$ and its variance estimator $\hat{V}(\hat{\mu}_y)$ at different sample sizes was investigated for the three different models. All models used the full set of covariates as predictors and no model selection was performed. For each iteration of the simulation, a sample of size n was taken from the 4151 observations. The different models were fit using the covariates in the sample and the percentage tree cover values were estimated as the response.

For each iteration of the simulation, $\hat{\mu}_y$ and $\hat{V}(\hat{\mu}_y)$ were calculated using equations (1), (2), and (3). This process was repeated for 1000 iterations and the mean of the variance estimate

$$E\left[\hat{V}(\hat{\mu}_y)\right] = \frac{1}{1000} * \sum_{i=1}^{1000} \hat{V}(\hat{\mu}_y)$$

was compared to the variance of the post-stratified estimator of the mean $Var(\hat{\mu}_y)$.

Results indicate that the EPSE estimator can be extended to include the Random Forests model. This simulation also supports the use of the EPSE estimator when the stratum boundaries are estimated quantiles from the model fits instead of fixed stratum boundaries. The justification for using the quantiles of the model predictions for the set of population covariates is to avoid having to deal with unequal sampling weights if present in the sampling design. The EPSE estimator appears to be robust for different models and under estimated stratum boundaries as long as care is taken to correctly specify the model and no optimization step is performed. There were also some technical issues involving numerical integration methods used in the statistical software not discussed in this paper. For this study, all elements in the population have equal probabilities of being sampled, but this method of using the predictions at the population level should allow for the use of an unequally weighted sampling design. This is an area for further research.

Properties of Minimization of EPSE Variance Estimates

For the second simulation the goal is to construct the smallest variance estimate $\hat{V}(\hat{\mu}_y)$ and to determine if this algorithm is performing well as an estimator for the true variance $Var(\hat{\mu}_y)$. The optimization will be over the number of strata to be used in the EPSE estimator. The model predictions will be split into stratum of equal size by the quantiles of the model predictions based on the covariates for the population.

We start this simulation as before by taking a random sample of size n from the population of 4151 sites in the Utah dataset. For each sample we fit a linear regression model, spline regression model, and Random Forests model using covariates to model percentage tree cover. Predictions were made for the sample values based on the model fits.

After computing the model predictions for the population, the strata over which optimization will be performed can be created. For $k = 1$, one stratum is used and is equivalent to simple random sampling. For arbitrary k, the predictions are placed into k equally spaced quantiles.

For a fixed n, the optimized EPSE estimator using Random Forests has the smallest variances and variance estimates of the three models, the spline model has the next smallest variances and variance estimates, and the linear model has the largest variances and variance estimates. The results also suggest that both the linear model and the spline model are overfitting the data resulting in the variance estimates being too small. This simulation study supports the use of Random Forests in EPSE, but suggests that an EPSE estimator based on linear model or spline model can underestimate the variance when an additional optimization is performed. This statement appears to hold for both the optimized and non-optimized EPSE estimates. This is a reasonable result since no model selection was performed for any of the models and therefore there are too many covariates for the spline model and the linear model.

CONCLUSION

This study has shown that use of the EPSE estimator should not be applied without a simulation study to determine if variance estimates for the EPSE estimator are over- or under-estimating the true variance. Furthermore, care must be taken in implementing the EPSE method with software that uses numerical integration methods to prevent rounding errors from influencing the results. Caution is needed when attempting to optimize the variance estimates as severe under-estimation of the variance of the EPSE estimator occurred in this study. The use of a model selection step in building the model has been shown to reduce this problem.

This study lends strength to the idea that EPSE can be applied to stratum boundaries that are estimated quanitles of the data rather than fixed stratum boundaries. This is an area for further research as in practice it is easier to implement the EPSE estimator using estimated quantiles and thereby eliminating the possibility of empty stratum. The Random Forests model performed well in each simulation and across all sample sizes considered. This is an exciting result in that there is almost no tuning needed by the user to fit the Random Forests model. This supports FIA's use of maps of land cover and percentage tree cover created by Random Forests as a basis for using endogenous post-stratification as a way to increase precision of FIA estimates.

ACKNOWLEDGMENTS

The authors would like to thank the FIA for support of this research. In addition, thanks go out to the FIA and the Remote Sensing Applications Center for the dataset used in this study.

LITERATURE CITED

Breidt, F.J.; Opsomer, J.D. 2008. **Endogenous post-stratification in surveys: classifying with a sample-fitted model.** The Annals of Statistics. 36(1): 403-427.

Breiman, L. 2001. **Random forests.** Machine Learning. 45: 5-32.

Coulston, J.W.; Moisen, G.G.; Wilson, B.T.; Finco, M.V.; Cohen, W.B.; Brewer, C.K. 2012. **Modeling percent tree canopy cover: a pilot study.** Photogrammetric Engineering and Remote Sensing. 78(7): 715-727.

Dahlke, M.; Breidt, F.J.; Opsomer, J.D.; Van Keilegom, I. [In press]. **Nonparametric endogenous post-stratification estimation.** Statistica Sinic.

Särndal, C.E.; Swensson, B.; Wretman, J. 2003. **Model assisted survey sampling.** New York, NY: Springer Verlag. 694 p.