OPTIMIZED ENDOGENOUS POST-STRATIFICATION IN FOREST INVENTORIES

Paul L. Patterson[1]

Abstract.—An example of endogenous post-stratification is the use of remote sensing data with a sample of ground data to build a logistic regression model to predict the probability that a plot is forested and using the predicted probabilities to form categories for post-stratification. An optimized endogenous post-stratified estimator of the proportion of forest has been recently proposed in the literature, but there are no known literature results describing the operating characteristics of this estimator. This study reports the results of a detailed Monte Carlo investigation of the performance of the optimized and another endogenous post-stratified estimator under a variety of realistic scenarios and compares their performance with earlier approaches.

INTRODUCTION

In recent years, estimators have been proposed that use remotely sensed data in conjunction with natural resource inventory sample data to construct land cover or use classifications and then use the classification to post-stratify the sample; Breidt and Opsomer (2008) proposed the term endogenous post-stratification (EPS) to describe the "post-stratification of the sample based on categories derived from the sample data." One example is using forest inventory sample data in conjunction with Landsat Thematic Mapper (TM) imagery to construct a logistic regression model that predicts for each pixel the probability the land associated with the pixel is forested, and then uses a set of strata boundaries so the land associated with the TM image can be stratified. An example of two strata would be as follows: any pixel with probability less than or equal to 0.5 is placed in one stratum, and any pixel with probability greater than 0.5 is in the other stratum. The stratification index (the predicted value from the logistic regression model) was derived using the sample, so the post-stratification is endogenous. Since endogenous post-stratification "violates the standard post-stratification assumptions that observations are classified without error into post-strata and the post-stratum population counts are known" (Breidt and Opsomer 2008), the statistical properties may differ. Breidt and Opsomer (2008) derived statistical properties of endogenous post-stratified estimators (EPSEs) in the case that the stratification index is the logistic regression model, and the stratum boundaries are determined independent of the sample data. McRoberts (2010) proposed an EPSE where the stratification index is a logistic regression model constructed from forest inventory sample data and TM imagery, and the strata boundaries are constructed based on an optimization procedure. To justify certain statistical properties of the optimized EPSE (OEPSE) McRoberts (2010) cites Breidt and Opsomer (2008), which is incorrect because the strata boundaries are derived from the sample data.

The objectives of this study are to (1) construct three EPSEs of the proportion of forest; (2) give a process for constructing populations that can be used in simulation studies to deduce the statistical properties of the three EPSEs; and (3) deduce the statistical properties of the three estimators based on simulation studies for three populations that represent a range of geographical regions and vegetative classes.

[1] Statistician, U.S. Forest Service, Rocky Mountain Research Station, 507 25th St., Ogden, UT 84401. To contact, call 907-295-5966 or email at plpatterson@fs.fed.us.

DATA AND ESTIMATORS

Three study areas were used to construct three simulation populations. The three study areas were: (1) the portion of path 27, row 27, Landsat scene in northern Minnesota, which is the study area in McRoberts (2010); (2) the portion of path 37, row 32, Landsat scene in northern Utah; and (3) path 37, row 33, Landsat scene totally within Utah (Fig. 1). These scenes, chosen because they represent differing geographical regions and vegetative classes, have an estimated forest cover of approximately 70, 50, and 30 percent respectively. TM imagery was acquired for three dates corresponding to early, peak, and late seasonal vegetative stages (Table 1). For each date, the normalized difference vegetative index and the tasseled cap transforms (brightness, greenness, and wetness) were used. Forest inventory data for permanent field plots established by the Forest Inventory and Analysis (FIA) Program were obtained for each of the scenes. This study used only the central subplot, which was associated with the image pixel that contains the center of the subplot. The numbers of totally forested, totally nonforested, and partially forested central subplots contained within each scene are in Table 1.

Following McRoberts (2010), the assumptions are (1) there is a finite population consisting of *N* elements which are 30 m × 30 m Landsat pixels; (2) there is an equal probability sample of *n* population units (3) with observed characteristic y_i, the proportion of forest for the land associated with each pixel; (4) the ancillary information for each population element, $\mathbf{x}_i$, is 12 Landsat-based spectral transforms; and (5) since the sample consists of FIA central subplots, then the central subplot is assumed to characterize the entire pixel that contains the center of the central subplot. The population parameter of interest is the mean proportion of forest, $\bar{Y} = \sum_i^N y_i / N$.

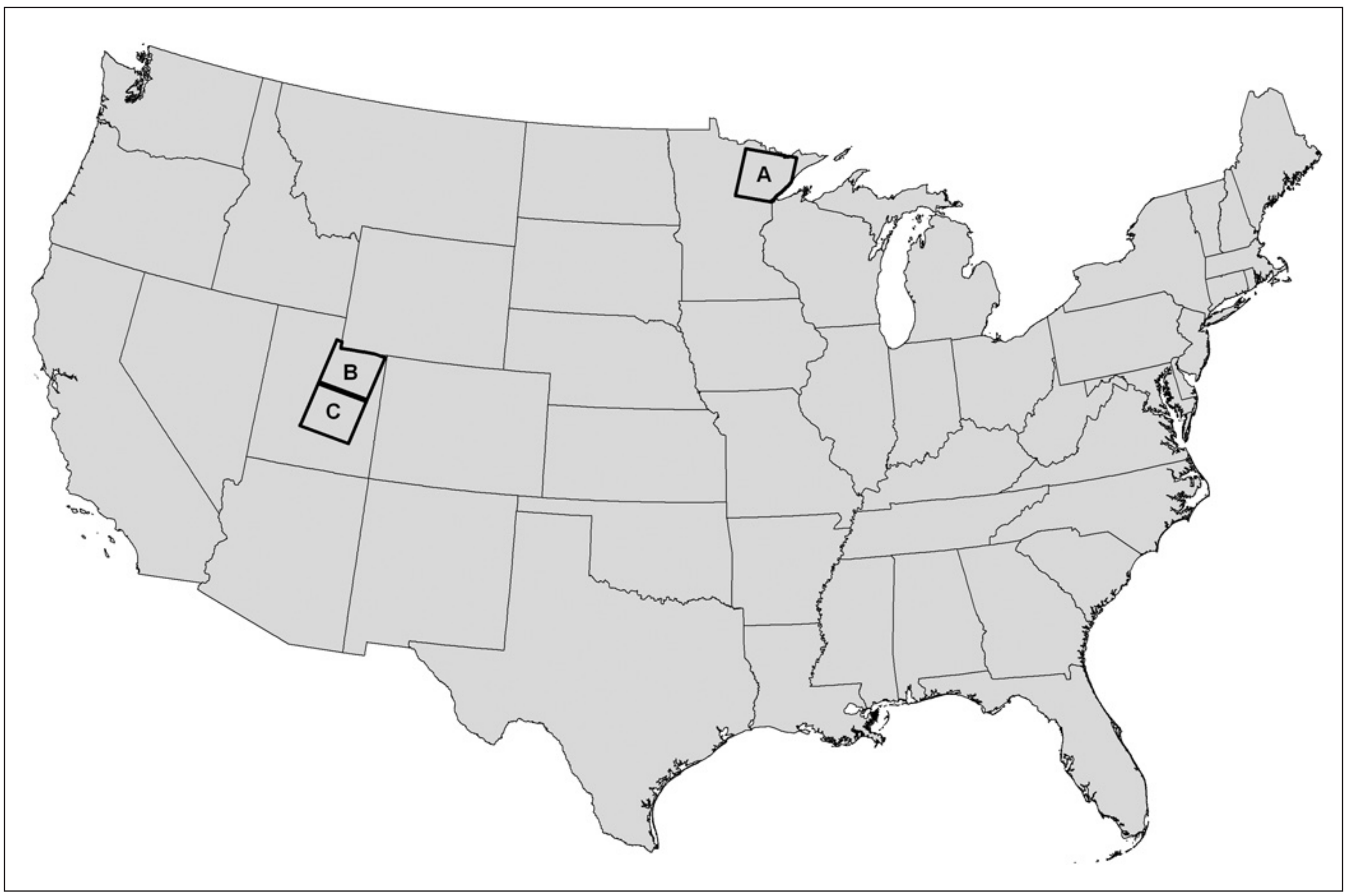

Figure 1.—Study areas: A is path 27, row 27; B is path 37, row 32; and C is path 37, row 33.

Table 1.—Dates for the three Landsat scenes, and number and type of FIA plots, used to construct the simulation populations. Dates represent early, peak, and late seasonal vegetative stages. Plots are those whose center subplot is located in the Landsat scene.

Remotely Sensed Data		Forest Inventory and Analysis Field Plots			
			Number of Plots		
Scene	Three Dates	Field Seasons	Forested	Nonforested	Partially Forested
Path 27, Row 27 within Minnesota	April 29, 2000 July 5, 2001 November 5, 1999	1999 to 2003	1,677	607	34
Path 37, Row 32 within Utah	June 14, 2000 July 30, 2002 October 10, 1999	2000 to 2009	659	620	15
Path 37, Row 33 in Utah	July 4, 2007 August 21, 2007 October 18, 2004	2000 to 2009	447	909	12

Breidt and Opsomer (2008) present post-stratification as a stratification index variable z_i, which along with a set of strata boundaries is used to partition the population into H strata; in endogenous post-stratification (EPS) the stratification index, strata boundaries, or both, are derived from the sample data. For example, the variable z_i is a forest/nonforest indicator, and is observed on a sample but is unknown for the rest of the population. The z_i are assumed to follow a logistic model, that is

$$\mathrm{E}(z_i|\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)} \qquad \mathrm{Var}(z_i|\mathbf{x}_i) = \nu, \qquad [1]$$

where the expectation is with respect to the model. The model parameters $\boldsymbol{\beta}$ are estimated by maximizing the likelihood using the subset of the sample for which the pixel is either completely forested ($\boldsymbol{y_i} = \mathbf{1}$) or completely nonforested ($\boldsymbol{y_i} = \mathbf{0}$). The estimate $\hat{\boldsymbol{\beta}}$ is used to define the endogenous stratification index $\hat{\boldsymbol{z}}_{\boldsymbol{i}} = \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)/\left(\mathbf{1}+\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)\right)$ for all population elements. To complete an EPS we need to specify the strata boundaries, which will be discussed in the following paragraph. Once the boundaries have been specified, the standard post-stratified estimator (PSE) $\bar{\boldsymbol{Y}} = \sum_i^{\boldsymbol{N}} \boldsymbol{y_i}/\boldsymbol{N}$ can be used, that is,

$$\hat{\mu}_{str} = \sum_{h=1}^{H} W_h \hat{\mu}_h$$

and [2]

$$\hat{V}(\hat{\mu}_{str}) = \left(1-\frac{n}{N}\right)\sum_{h=1}^{H} W_h^2 \frac{\hat{\sigma}_h^2}{n_h},$$

where $\hat{\mu}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$, $\hat{\sigma}_h^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\hat{\mu}_h)^2$, W_h is the weight of stratum h, y_{hi} is the proportion of forest for the ith sample element of stratum h, and $\hat{V}(\hat{\mu}_{str})$ is an approximate estimated variance for the PSE.

Three methods for specifying strata boundaries will be presented and used along with $\hat{\mu}_{str}$ from Equation [2] to define an EPSE of the mean proportion of forest. All three schemes use the endogenous stratification index, $\hat{z}_i$, defined above. The first EPS scheme uses fixed predetermined strata boundaries, $0 = \tau_0 < \tau_1 < \ldots < \tau_{H-1} < \tau_H = 1$; strata assignment given by $\hat{z}_i$ will be denoted as the fixed boundary endogenous post-stratified estimator (FEPSE). The second EPS scheme specifies the strata boundaries using an optimization process. For strata boundaries $\{\tau_h\}_{h=1}^{H-1}$ with values rounded to the nearest hundredth and for which each stratum contains at least four sample elements, calculate $\hat{V}(\hat{\mu}_{str})$ (Equation [2]). Determine the set of boundaries, over all possible numbers of strata, which minimizes $\hat{V}(\hat{\mu}_{str})$; the OEPSE is calculated using this set of strata boundaries. McRoberts (2010) found little reduction in $\hat{V}(\hat{\mu}_{str})$ when optimizing over two or three strata compared to only two strata. Two strata are defined by a single stratum boundary; a natural choice for that boundary would be the mean proportion of forest, which is estimated by the sample mean, that is, $\hat{\tau} = n^{-1}\sum_{i=1}^{n} y_i$. The third EPSE, which is denoted the estimated endogenous post-stratified estimator (EEPSE), is calculated using the strata boundaries $\{0, \hat{\tau}, 1\}$.

In EPS the classification of observations and the post-stratum population counts depend on the sample and these two aspects add additional sources of variability. Breidt and Opsomer (2008) showed for a class of EPSEs that $\hat{V}(\hat{\mu}_{str})$ converges to the asymptotic variance of the EPSE, as $n, N \rightarrow \infty$. Breidt and Opsomer's assumptions that are germane to our study are: first, the set of stratification indices they consider contains $\hat{z}_i$ defined above; and second, the strata boundaries are fixed. Hence their result on the asymptotic variance applies to the FEPSE, but not to the OEPSE or the EEPSE.

SIMULATION POPULATION AND SIMULATION STUDY

For each of the three Landsat scenes, the population is the landmass covered by the scene; population elements are the land delineated by the image pixels. The FIA reduced sample refers to the set of FIA central subplots which are contained in the scene and are completely forested or completely nonforested. The goal is assign a value of forest or nonforest to each pixel so that the logistic relationship between the values of the 12 Landsat-based spectral transforms and the forest/nonforest values for the entire scene is similar to the logistic relationship between the observed forest/nonforest values for the FIA reduced sample and the value of 12 Landsat-based spectral transforms for pixels associated with the FIA reduced sample.

First, the FIA reduced sample and the associated 12 Landsat-based spectral transforms, $\mathbf{x}_i$, were used to estimate the parameters $\boldsymbol{\beta}$ of the logistic regression model, Equation [1]. Denote the fitted model by $m(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)$ and for $0 < \tau < 1$ define the variable $v_{i\tau} = \begin{cases} 0 & \text{if } m(\boldsymbol{\beta}'\mathbf{x}_i) \le \tau \\ 1 & \text{if } m(\boldsymbol{\beta}'\mathbf{x}_i) \le \tau \end{cases}$, which is a forest/nonforest indicator based on the predicted probability of forest from the logistic model and the cutoff τ. For the FIA reduced sample a measure of misfit between $v_{i\tau}$ and the proportion of forest, y_i, was defined as $d_\tau = n^{-1}\sum_{i=1}^{n}|y_i - v_{i\tau}|$, where n is the size of the FIA reduced sample. The measure of misfit d_τ was minimized over the grid $\tau \in \{0.05, 0.10, \ldots, 0.95\}$; denote where the minimum occurred by τ_0. Then a forest/nonforest indicator variable, z_i, was defined for every pixel in the scene so the measure of misfit between z_i and $v_{i\tau_0}$ over the population was the same as the measure of misfit between y_i and $v_{i\tau_0}$ over the FIA reduced sample. For the simulation population the proportion of forest is $\bar{Z} = \sum_i^N z_i / N$.

From the simulation population 4,000 simple random samples of size n were drawn. For each sample, j, four estimates of the proportion of forest were calculated: the simple random sample estimate $\left(\hat{\mu}_{SRSj} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}\right)$, the FEPSE with boundaries $\{0, 0.5, 1\}$, the OEPSE, and the EEPSE. The estimates for each sample are denoted by $\hat{\mu}_{*j}$, and the value of the proposed estimated variance is denoted by $\hat{V}(\hat{\mu}_{*j})$, where the $*$ is the acronym for the estimator. To determine both the small sample size properties and the asymptotic behavior, sample sizes of 100, 200, … , 700 were used (fitting the logistic model was problematic at smaller sample sizes). For each sample size the statistical properties were evaluated using (1) the empirical bias, EBias($*$); (2) the empirical mean squared error, EMSE($*$); and (3) the empirical variance, EV(*), where $*$ indicates the estimator.

To measure whether the proposed estimated variances of the three EPSE converge to the variance of the estimator, the actual coverage for 95-percent confidence intervals was calculated, that is, the percentage of the 4,000 replicates where the interval $\mu_{*j} \pm 1.96\{\hat{V}(\hat{\mu}_{*j})\}^{1/2}$ contains the proportion of forest $\bar{Z}$. A standard measure used to compare two unbiased estimators is the relative efficiency between the two estimators. For example, the relative efficiency (RE) between EEPSE and OEPSE is: RE(EEPSE, OEPSE) = (EMSE(EEPSE)) / (EMSE(OEPSE)).

RESULTS AND DISCUSSION

All the EPSEs were empirically unbiased at all sample levels. Hence the variance can be used in lieu of the mean squared error. The acceptance region for n trials of the empirical coverage percentages is $p \pm z_{1-\alpha}\sqrt{p(1-p)/n}$, where p is the stated confidence level, n is the number of simulations, and α is the confidence level for the acceptance region. Figures 2 and 3 show the empirical coverage percentages for two simulation populations; the third was similar. For all three scenes the pattern of convergence is the same: the EEPSE converges at the fastest rate, followed by the FEPSE and then the OEPSE. For sample size 100, the OEPSE empirical coverage percentage is well below the empirical coverage percentages of the other two EPSEs, indicating $\hat{V}(\hat{\mu}_{OEPSEj})$ significantly underestimates the true variance. All the empirical coverage percentages were in the acceptance region when the sample sizes were approximately equal to the number of FIA plots in the scene.

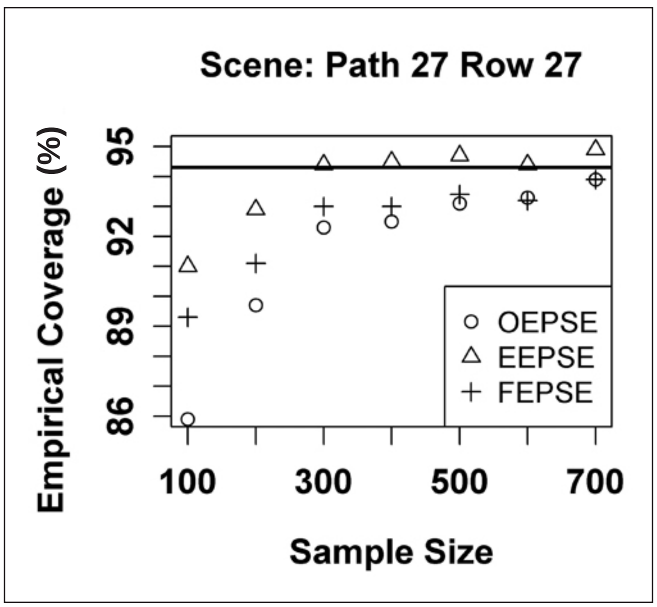

Figure 2.—For the simulation population path 27, row 27 the empirical coverage percentages of the nominal 95-percent confidence intervals for the optimized endogenous post-stratified estimator (OEPSE), the fixed boundary endogenous post-stratified estimator (FEPSE), and the estimated endogenous post-stratified estimator (EEPSE). The empirical coverage percentage is based on 4,000 realizations. The horizontal line y=94.3 is the lower bound for the acceptance region at confidence level of 95 percent.

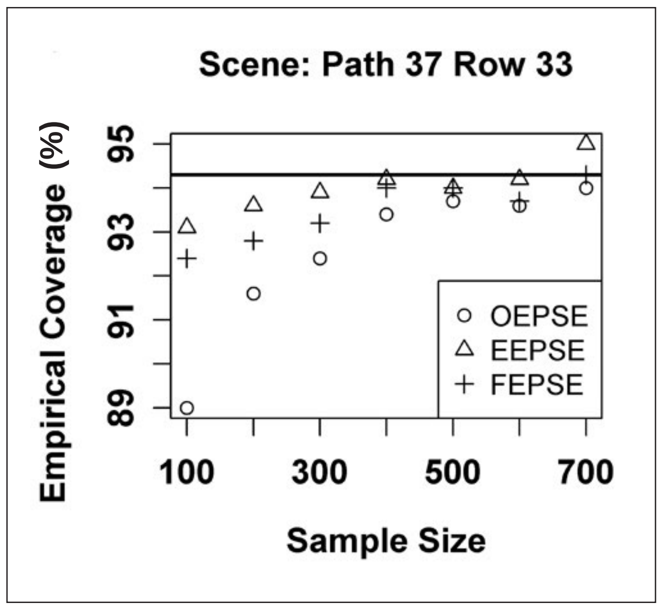

Figure 3.—For the simulation population path 37, row 33 the empirical coverage percentages of the nominal 95-percent confidence intervals for the optimized endogenous post-stratified estimator (OEPSE), the fixed boundary endogenous post-stratified estimator (FEPSE), and the estimated endogenous post-stratified estimator (EEPSE). The empirical coverage percentage is based on 4,000 realizations. The horizontal line y=94.3 is the lower bound for the acceptance region at confidence level of 95 percent.

The pattern for relative efficiencies RE(EEPSE,OEPSE) and RE(SRS,OEPSE) is the same for all three simulation populations: the empirical relative efficiency increases as the sample size increases. Table 2 contains the relative efficiency for sample sizes of 100, either 500 or 700, and the sample size approximately equal to the number of FIA plots. The results indicate the EEPSE appears to be uniformly more efficient than the simple random sample estimator (SRS). The rest of the results paint a mixed picture. In the arid Interior West (Path 37, Rows 32 and 33) the EEPSE is more efficient than the OEPSE at small sizes and slightly less efficient for the larger sample sizes; the EPSEs appear to have around 30-percent improvement in efficiency over the SRS for large sample sizes and no improvement for small sample sizes. In the heavily forested Upper Midwest (Path 27, Row 27) the OEPSE appears to have a gain in efficiency over the EEPSE for large sample sizes and a 50- to 60-percent improvement over SRS.

This study's results are applicable to situations where most if not all of the sample values for the proportion of forest are either 0 or 1. Four conclusions can be drawn. First, the three endogenous post-stratified estimators appear to be unbiased for sample sizes 100 and greater. Second, the standard estimated variance for post-stratified estimators appears to asymptotically converge to an estimate of the true variance for the estimated and optimized endogenous post-stratified estimators. Third, the estimated endogenous post-stratified estimator's asymptotic variance appears to have the faster rate of convergence, followed by the fixed boundary estimator and then the optimized estimator. Fourth, the estimated and optimized endogenous post-stratified estimators appear to have a higher efficiency compared to the simple random sample estimator. Additionally, it appears that for some land cover and large sample sizes the optimized endogenous post-stratified estimator has greater efficiency than the estimated endogenous post-stratified estimator.

LITERATURE CITED

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

McRoberts, R.E. 2010. **Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data.** Remote Sensing of Environment. 114: 1017-1025.

Table 2.—The empirical relative efficiencies (RE) between the estimated endogenous post-stratified estimator (EEPSE), the optimized endogenous post-stratified estimator (OEPSE), and the simple random sample estimator (SRS), for three sample sizes for each of the three simulation populations. The empirical relative efficiencies are based on 4,000 realizations.

	Simulation Population								
	Path 27 Row 27			Path 37 Row 32			Path 37 Row 33		
Sample Size	100	700	2,280	100	500	1,270	100	700	1,350
RE(EEPSE,OEPSE)	1.01	1.15	1.19	0.90	0.97	1.00	0.92	1.03	1.06
RE(SRS,OEPSE)	1.11	1.53	1.59	0.98	1.26	1.35	0.99	1.22	1.28
RE(SRS,EEPSE)	1.10	1.33	1.34	1.09	1.30	1.35	1.08	1.18	1.21