ESTIMATORS USED IN THE NEW MEXICO INVENTORY: PRACTICAL IMPLICATIONS OF "TRULY" RANDOM NONRESPONSE WITHIN EACH STRATUM

Paul L. Patterson and Sara A. Goeking[1]

Abstract.—The annual forest inventory of New Mexico began as an accelerated inventory, and 8 of the 10 Phase 2 panels were sampled between 2008 and 2011. The inventory includes a large proportion of nonresponse. FIA's estimation process uses post-stratification and assumes that nonresponse occurs at random within each stratum. We construct an estimator for the New Mexico inventory and derive an estimated variance based on the missing-at-random assumption.

INTRODUCTION

The national Forest Inventory and Analysis (FIA) Phase 2 grid forms the basis for FIA sampling, yet not all Phase 2 plots can be sampled. When plots are not sampled due to denial of access, logistical constraints, or hazardous conditions, we refer to them as nonresponse following the convention of Patterson et al. (2012). From 2008 to 2011 in New Mexico 8 of 10 panels (more than 5,000 plots) were sampled under an accelerated inventory. For many reasons, there is a large amount of nonresponse in New Mexico, which warrants an examination of how FIA handles nonresponse.

Standard FIA assumptions are that we have a simple random sample of a region, *R*, which can be post-stratified and whose total area, A_T, is known. The number of acres, A_d, in domain *d* is of interest, and is equal to $A_T P_d$, where P_d is the proportion of *R* that is classified in domain *d*. The FIA post-stratified estimator for P_d is

$$\hat{P}_d = \Sigma_{h=1}^{H} W_h \hat{P}_{dh} \qquad [1]$$

where *H* is the number of strata, W_h is the weight of the strata, and $\hat{P}_{dh}$ is an estimate of the proportion of stratum *h* that is in domain *d*. The definition of $\hat{P}_{dh}$ in chapter 4 of Bechtold and Patterson (2005) contains adjustments for nonresponse plots. The two types of nonresponse are entire plots and partial plots. The adjustment for plots that are entirely nonresponse is to reduce the sample size, while an adjustment factor is used to compensate for plots with partial nonresponse. Both adjustments are based on the assumption that nonresponse is random within each stratum. Patterson et al. (2012) showed by example that substantial bias can occur if the missing-at-random assumption is violated in the case where the nonresponse is restricted to entire plots. Roesch et al. (2012) showed a) why bias occurs when there are plots with partial nonresponse, and b) that if the missing-at-random assumption is correct, then the adjustment factor for the partial nonresponse portion of the estimator is unbiased, given the assumption the sample size is reduced to adjust for plots that are entirely nonresponse. The issue of whether the reduction in sample size, through a non-random process, induces a bias in the estimator is not addressed. Neither is the effect of the reduction in sample size, through a non-random process, on the variance and estimated variance addressed in Bechtold and Patterson (2005) or Roesch et al. (2012).

[1] Statistician (PLP) and Biological Scientist (SAG), U.S. Forest Service, Rocky Mountain Research Station, 507 25th St., Ogden, UT 84401. PLP is corresponding author: to contact, call 907-295-5966 or email at plpatterson@fs.fed.us.

Roesch et al. (2012) proposed using partitions of the strata, for which the assumption that the nonresponse is random is tenable for each of the partitions, and then for each partition use ($\hat{P}_{d*}$) to estimate the portion of the partition that is in the domain d. In Särndal et al. (1992) the division of the population into groups where the assumption of random nonresponse is valid is referred to as the response homogeneity group (RHG) model.

The purpose of this manuscript is to: 1) construct the estimator used to estimate P_d for the estimation units in the New Mexico inventory; 2) use the statistical framework in Särndal et al. (1992) to construct an estimator equal to the estimator constructed in item 1 and use this equality to investigate the statistical properties of the FIA estimator, namely bias, variance and properties of an estimated variance; and finally 3) discuss the implications for the current FIA estimated variance.

NEW MEXICO ESTIMATOR

Two potential partitioning criteria arise in many FIA estimation units (Patterson et al. 2012). The first is ownership class; denied access is higher among private owners. The second is an aspect of the FIA pre-field procedure, in which high-resolution photos are used in conjunction with old field notes to classify each plot as either "office plot" or "field plot". The office plots are designated as nonforested. The salient point here is that for the office plots the probability of nonresponse is zero, while in most field plots the probability of nonresponse is greater than zero. Whether a stratum needs to be partitioned into office and field domains depends on two factors: first, the agreement between the stratification scheme and pre-field classification and second, the amount of nonresponse in the stratum. If a stratum needs to be partitioned into office and field substrata, the weights for these two partitions must be estimated.

To facilitate the comparison with Särndal et al. (1992) we will develop some notation. It is useful to differentiate between the strata-partitions where the weight is known and partitions where the weight must be estimated. We assume there are H strata with known weights. Of the H strata, H_1 do not need to be partitioned further and H_2 of the strata need to be partitioned further with the weights for the partitions being estimated. Note that $H_1 + H_2$= H. Let J_h denote the number of partitions in the hth stratum (note: for the first H_1 strata $J_h = 1$). In RHG model terminology (Särndal 1992) there are $J = \sum_{h=1}^{H} J_h$ response groups and each response group is a subset of one of the H strata.

In FIA's Interior West region there are typically two strata, Green (G) and Brown (B), representing forest and nonforest, with weights WG and WB, respectively. Because the nonresponse rate for the New Mexico annual inventory is unusually high on private lands, further stratification into partitions of Private owners (P) and Non-Private owners (NP) must be considered. The four strata weights are W_{GP}, W_{GNP}, W_{BP}, and W_{BNP}, with $W_{GP} + W_{GNP} + W_{BP} + W_{BNP} = 1$, where, for example, W_{GP} is the weight for the stratum which is green and private ownership (see Table 1). Nonresponse rates have been calculated for each stratum based on preliminary data from 2008 to 2011. Nonresponse rates for Private owners within the G and B strata are 47 percent and 19 percent, respectively, while rates for the Non-Private owners are 8 percent and 3 percent. The nonresponse rate for the NP partition is similar among all ownership subclasses (e.g., National Forest system, Bureau of Land Management). In addition, the nonresponse rates for both GNP and BNP are small enough that we can ignore the slight bias caused by blending the field plots and the office plots.

Table 1.—Calculated values of stratum weights W_{GP}, W_{GNP}, W_{BP}, and W_{BNP} for New Mexico, based on spatial intersection of the green/brown stratification and a statewide ownership layer

	Partition	
Stratum	Private (P)	Non-Private (NP)
Green (G)	0.057	0.155
Brown (B)	0.382	0.406

In contrast, the high nonresponse rates for both GP and BP warrant investigation of further partitioning into “field” (F) and “office” (O), based on pre-field determinations. Preliminary data from New Mexico indicate that nearly all plots in GP stratum were field visits, so there is no further partition of the GP stratum while approximately 40 percent and 60 percent of the BP stratum was field visit and office respectively. Let $w_{F(BP)}$ indicate the estimated weight (or proportion) of F partition within the BP stratum with similar definitions for weights of the O partition of the BP stratum. In terms of the H notation, $H = 4$, $H_1 = 3$ and $H_2 = 1$. The estimate is

$$\hat{P}_d = W_{GNP}\hat{P}^o_{d|GNP} + W_{BNP}\hat{P}^o_{d|BNP} + W_{GP}\hat{P}^o_{d|GP} + W_{BP}\left(w_{F(BP)}\hat{P}^o_{d|F(BP)} + w_{O(BP)}\hat{P}^o_{d|O(BP)}\right)$$

where $\hat{P}^o_{d|*}$ is the proportion of the $*$ stratum or partition that is in domain d and the superscript o indicates the proportion is based on the observed values and adjusted for the nonresponse in the stratum or partition. The explicit formula for $\hat{P}^o_{d|*}$ is given next.

We will now return to constructing the estimator, using notation consistent with that in Särndal et al. (1992). For the hth stratum denote the total number of plots by n_h and the number of partly or fully observed plots by m_h, so the number of entirely nonresponse plots is $n_h – m_h$. In FIA m_h is denoted by n_h (Bechtold and Patterson 2005) and there is no notation for the total number of plots. For the jth partition of the hth stratum let n_{hj} denote the total number of plots and m_{hj} denote the number of partly or fully observed plots. So $\sum_j^{J_h} n_{hj} = n_h$ and $\sum_j^{J_h} m_{hj} = m_h$. The estimated weight of jth partition of the hth stratum is $w_{hj} = \frac{n_{hj}}{n_h}$, where if $h = 1, \dots , H_1$, then $J_h = 1$ and $w_{h1} = 1$.

The estimate of the proportion of the partition hj that is in domain d is

$$\hat{P}^o_{d|hj} = \frac{\sum_{i=1}^{m_{hj}} a^d_{hji}}{\sum_{i=1}^{m_{hj}} a^o_{hji}} \qquad [2]$$

where a^d_{hji} is the area of observable land in domain d on the ith plot of the jth partition of the hth stratum, and a^o_{hji} is the amount of observable land on the ith plot of the jth partition of the hth stratum. For the H_1 strata that are not further partitioned, this equation reduces to

$$\hat{P}^o_{d|h} = \frac{\sum_{i=1}^{m_h} a^d_{hi}}{\sum_{i=1}^{m_h} a^o_{hi}} \qquad [3]$$

Two points are noteworthy. First, we are using the superscript notation used in Roesch et al. (2012) instead of the notation used in Bechtold and Patterson (2005). Second, equations [2] and [3] are the same as equation [4.1] in Bechtold and Patterson (2005) (with m_h substituted for n_h), where we have simplified equation [4.1] by cancelling terms. Combining all these parts, we have

$$\hat{P}_d = \sum_{h=1}^{H_1} W_h \hat{P}^o_{d|h} + \sum_{h=1}^{H_2} W_h \sum_{j=1}^{J_h} w_{hj}\hat{P}^o_{d|hj} \qquad [4]$$

The denominator in equations [2] and [3] is the adjustment for the partial nonresponse plots.

STATISTICAL PROPERTIES OF ESTIMATOR

We will be using the statistical model presented in section 15.6 of Särndal et al. (1992), which assumes each sample is decomposed into response groups and probability that an element in the kth group responds is a constant. Other technical assumptions are not stated here. The estimator of interest is constructed using the response groups and auxiliary variables (Särndal et al. 1992: section 15.6.4); the auxiliary variable in our case is strata membership. Two assumptions related to FIA merit further discussion.

First, if the adjustment factor is treated as a random quantity, then we have the combination of a ratio estimator and a response model, and the derivation of the statistical properties of the estimator becomes a much more onerous task. However, preliminary data for New Mexico indicate the number of plots with partial nonresponse is less than 1 percent, so treating the adjustment factor as fixed (a constant) rather than random has little effect on the variance.

Second, in section 15.6 of Särndal et al. (1992) the goal is to estimate the population total. Estimates of the total can be adjusted to proportions by dividing by N, the number of population elements. In using Särndal et al.'s results, we are approximating an infinite population with a finite population (Roesch et al. 2012).

Using these two assumptions, the estimator given by Equation 15.6.16 in Särndal et al. (1992) is equal to $\hat{P}_d$ in Equation [4]. From Result 15.6.2 of Särndal et al. (1992) $\hat{P}_d$ is approximately unbiased and an approximate variance is given. Of interest to us is the formula for a variance estimator. For a plot let the observed sample value be $y_* = a_*^d / \sum_{i=1}^{m_\#} a_{\#i}^o$ where $*$ indicates the index of the plot and # indicates the stratum/partition the plot is in, so the denominator is the adjustment factor for partial nonresponse plots. Let $r_\#$ denote the sample plots in stratum/partition # with either partial or total response. Then the estimated variance is given below. For clarity we have split the equation into three parts; the first part is for strata with $J_h = 1$ and the following two parts for strata with $J_h > 1$.

$$\hat{V}(\hat{P}_d) = \sum_{h=1}^{H_1} \left(w_h^2 (m_h^{-1} - n_h^{-1}) + w_h \frac{1-f}{n} (1 - \delta_h) \right) S_{yr_h}^2 + \quad [5]$$

$$\sum_{h=1}^{H_2} \left(w_h^2 \sum_{j=1}^{J_h} w_{hj}^2 (m_h^{-1} - n_h^{-1}) S_{yr_{hj}}^2 + \right. \quad [6]$$

$$\left. w_h \frac{1-f}{n} \sum_{j=1}^{J_h} w_{hj} (1 - \delta_{hj}) S_{yr_{hj}}^2 \right) +$$

$$\frac{n}{n-1} \sum_{h=1}^{H_2} w_h \sum_{j=1}^{J_h} w_{hj} \bar{e}_{r_{hj}}^2 \quad [7]$$

where $\delta_\# = 1 - \frac{1 - n_\#/n}{m_\#} \frac{n}{n-1}$, with # indicating the stratum/partition, is an adjustment factor for nonresponse; $S_{yr_\#}^2$ is the variance of the y_k for the response portion of the sample contained in stratum/partition #; and $\overline{e}_{r_{h*}}$ is the average of the residuals, $e_k = y_k - \sum_{j=1}^{J_h} w_{hj} \sum_{i=1}^{m_{hj}} y_{hi}$, over r_{h*}.

DISCUSSION

If $H_2 = 0$, then equation [4] is the current FIA estimator. The estimated variance we are proposing would involve only equation [5], which differs from the estimated variance FIA currently uses. The difference between our estimated variance and the one used by FIA is that the latter is derived based on the assumption that a simple random sample of size $m = \sum_h m_h$ is drawn, while our estimated variance is based on two-phase Bernoulli sampling for stratification, that is, a simple random sample s_a is drawn at the first phase and s_a is stratified into the response groups. Then a subsample s_{ah} is drawn from each s_a using Bernoulli sampling, with probability of selection equal to the response probability. This new approach may improve FIA's estimates in situations with high nonresponse where the assumption of missing at random within strata is untenable.

LITERATURE CITED

Bechtold, W.A.; Patterson P.L., eds. 2005. **The enhanced Forest Inventory and Analysis program-national sampling design and estimation procedures.** Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p.

Patterson, P.L.; Coulston, J.W.; Roesch, F.A.; Westfall, J.A.; Hill, A.D. 2012. **A primer of nonresponse in the U.S. Forest Inventory and Analysis program.** Environmental Monitoring and Assessment 184(3): 1423-1433.

Roesch, F.A.; Coulston, J.W.; Hill, A.D. 2012. **Statistical properties of alternative national forest inventory area estimators.** Forest Science. doi.org/10.5849/forsci11-008. Available at http://masetto.ingentaselect.co.uk/fstemp/b12e1416 7fb0e933f0f6fefca6a3f3c1.pdf. [Date accessed unknown].

Särndal, C.; Swensson, B.; Wretman, J. 1992. **Model assisted survey sampling.** New York: Springer-Verlag. 694 p.

The content of this paper reflects the views of the author(s), who are responsible for the facts and accuracy of the information presented herein.