Comparing Mapped Plot Estimators

Paul C. Van Deusen[1]

Abstract.—Two alternative derivations of estimators for mean and variance from mapped plots are compared by considering the models that support the estimators and by simulation. It turns out that both models lead to the same estimator for the mean but lead to very different variance estimators. The variance estimators based on the least valid model assumptions are shown to perform poorly in practice.

Introduction

Mapped plots are an integral part of the U.S. Department of Agriculture Forest Service Forest Inventory and Analysis (FIA) annual forest inventory design. The concept of mapping is intended to reduce the potential bias of having more than one forest condition on a single plot. The plot is mapped to indicate what proportion of the plot is covered by a particular condition. In the past, plots were moved or rotated into a uniform condition to avoid this problem. Plot rotation leads to a small bias in the estimates (Birdsey 1995), and mapping was believed to be a more statistically defensible approach (Hahn *et al.* 1995).

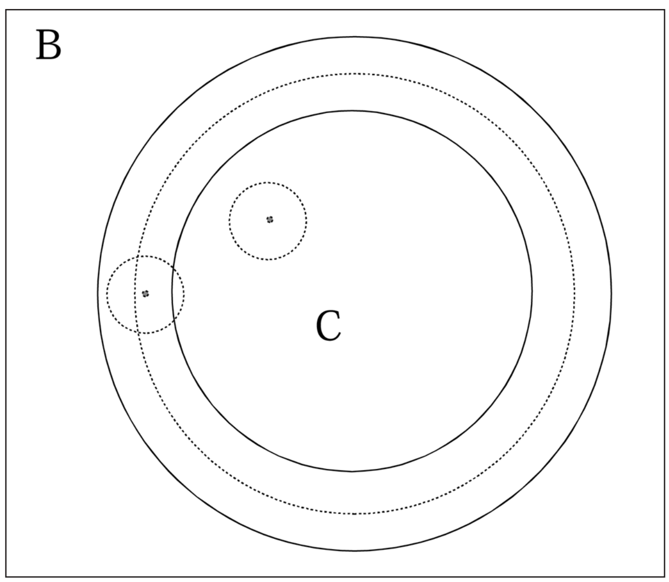
More than one estimator has been suggested for obtaining estimates of means and variances from mapped plots. Estimator 1 (EST1) is described in a FIA document (FIA 2004), and estimator 2 (EST2) is described in Van Deusen (2004). These estimators will be derived using a model-based approach and then compared.

Review of Theory

The FIA plot design consists of four circular subplots in a fixed configuration. Small diameter trees are measured only on smaller concentric plots in the larger subplots. Ignoring the specifics of the FIA plot design simplifies derivation of the estimators without loss of generality. Therefore, the estimators are derived under the assumption that the sample plots consist of a single fixed-area plot.

The simple forest inventory model used in Van Deusen (2004) is followed here. Assume two conditions exist, C and B, where C is a circular condition surrounded by condition B (fig. 1). Estimates of the mean and variance of type C are of interest, and the other types that surround it are denoted as B. The shape of type C could be anything, in practice. Sampling is either systematic or simple random and uses fixed-area circular plots with radius d. The edge of type C is shown by a dash line, and a perimeter band that overlaps the outer edge of area C is shown by solid lines. The plot contains both conditions when the plot center falls within the perimeter band. The condition boundary is mapped when it crosses a plot.

Figure 1.—*An area of condition C surrounded by condition B. Condition C is bounded by the dash line that is contained in a perimeter band of width equal to the diameter of the fixed-area circular plots. Plots with centers that fall within the band will contain some of both conditions. All other plots contain only one condition. One plot that is fully in condition C is shown along with a plot that overlaps the boundary. Plots that contain no C are not of interest.*



[1] National Council for Air and Stream Improvement, 600 Suffolk Street, 5th floor, Lowell, MA 01854. E-mail: pvandeusen@ncasi.org.

The following notation is similar to that used in Van Deusen (2004):

a_i = The proportion of the area of plot i that is within condition C.

n = The number of plots that contain some condition C.

$$\bar{a} = \sum_{i=1}^{n} a_i / n \ .$$

Where:

y'_i = A variable that can be measured on each randomly located plot that completely or partially overlaps condition C. For plots that don't overlap C, $y'_i = 0$.

$y_i = y'_i/p_i$ is the original measurement expanded to a per acre (hectare) or total value.

p_i = A value proportional to the selection probability of the plot, e.g., this would be 1/5 for a fifth-acre circular plot.

μ = The per unit area mean of variable y for condition C, e.g., cubic meter per hectare pine volume.

After a plot is located, the amount of variable y' is recorded and expanded to a per acre value, y. When plot i contains none of condition C, $y_i = 0$, and a_i=0. The n plots that contain a non-zero amount of condition C are labeled 1,...,n.

Estimator Derivation

A model-based approach is used to derive mapped plot estimators. The estimators called EST1 and EST2 in the introduction are both derived in this section.

EST1 is derived first by starting with the following model:

$$y_i = \bar{a}\,\mu + e1_i \qquad (1)$$

where e1 is a random error term and $E(e1_i, e1'_i) = \bar{a}^2 \sigma^2$. This model states that the expanded plot value is equal to the mean per unit area value, μ, multiplied by the average proportion of a plot, $\bar{a}$, that is in the condition of interest. The variance of the expanded plot value is proportional to the squared average proportion in the condition.

Consider the result of dividing equation (1) by $\bar{a}$ to get the following equation:

$$y_i / \bar{a} = \mu + e_i \qquad (2)$$

where $E(e_i, e'_i) = \sigma^2$. The transformed variable, $\tilde{y} = y_i / \bar{a}$, is the variable that the estimators in the "Stat Band" document (FIA 2004) are based on. Thus, the following estimators for mean and variance are based on this equation:

$$\bar{Y} = \frac{\sum_{i=1}^{n} \tilde{y}_i}{n} \text{, and} \qquad (3a)$$

$$v_1(\bar{Y}) = \frac{\sum_{i=1}^{n} \tilde{y}_i^2 - n\bar{Y}^2}{n(n-1)} \qquad (3b)$$

These are standard mean and variance estimators that would apply to simple random or stratified random sampling without a finite population correction factor (Cochran 1977, Thompson 2002).

EST2 is derived from a related model that allows for plot values to vary according to the percentage of the plot falling in the condition:

$$y_i = a_i\,\mu + e2_i \qquad (4)$$

where $E(e2_i, e2'_i) = a_i \sigma^2$. Equation (4) states that the amount, y_i, in the condition is related to the proportion, a_i, of the plot that falls in the condition. The variance of y_i is also proportional to a_i, which seems intuitively reasonable. Using standard least-squares formulas, the following are the estimators for the mean and variance (Van Deusen 2004):

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} a_i} \text{, and} \qquad (5a)$$

$$v_2(\bar{Y}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} a_i} \text{, where} \qquad (5b)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - a_i \bar{Y})^2}{\sum_{i=1}^{n} a_i - 1} \quad (5c)$$

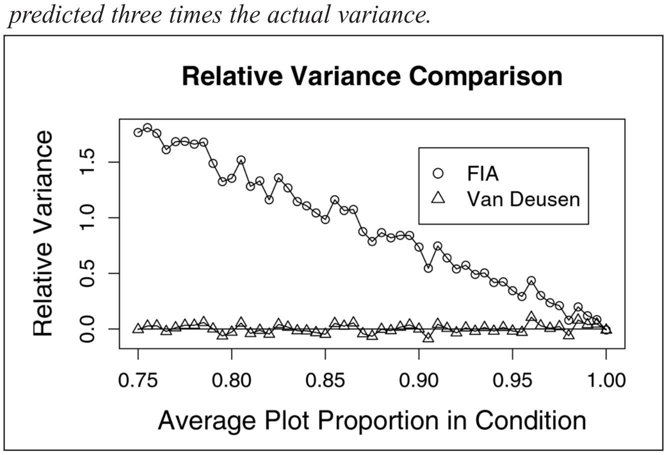
More detailed discussion about the derivation of equations (5b and 5c) can be found in Van Deusen (2004).

Note that equations (3a) and (5a) are identical because $n\bar{a} = \sum a_i$. The variance estimators, however, are quite different. Intuition suggests that v_2 should be less than v_1 because equation (4) is more flexible than equation (1). The exception is when all plots are fully in the condition so that $\sum a_i = n$, making equations (3b) and (5b) identical. In general, equation (1) is a crude model of the relationship between the plot values, y_i, and the proportions, a_i.

Simulated Comparison

Simulated data are used to compare variance estimators v_1 and v_2. The same simulation scheme used in Van Deusen (2004) is used here. Y-values are drawn from a normal distribution with standard deviation 300 and mean 1,000. There are 1,000 replications with 100 samples per replication. The average proportion of the plot in the condition area varies from 0.75 to 1.0. More simulation details are available in Van Deusen (2004).

Figure 2.—*Simulated comparison of two estimators for variance of the mean from mapped plots. A relative variance of 0 means the variance estimator was unbiased, a value of 1 means it predicted twice the actual variance, and a value of 2 means it predicted three times the actual variance.*



The simulation results in no noticeable bias in estimates of the mean (Van Deusen 2004), and the variance estimates are identical when $\bar{a} = 1.0$, as they should be. Variance estimator v_1 deteriorates, however, as $\bar{a}$ moves away from 1.0 (fig. 2). The relative variances (fig. 2) are computed as $(\hat{v} - v)/v$, where $\hat{v}$ is the average of the estimated variances based on 100 observations from each of the 1,000 replications, and v is the simulated variance computed from the means of the 1,000 replications. A relative variance of 0 means the formula is unbiased, a value of 1 shows that the formula is predicting twice the variance that it should, and a value of 2 indicates that the formula is overpredicting by a factor of 3.

Discussion

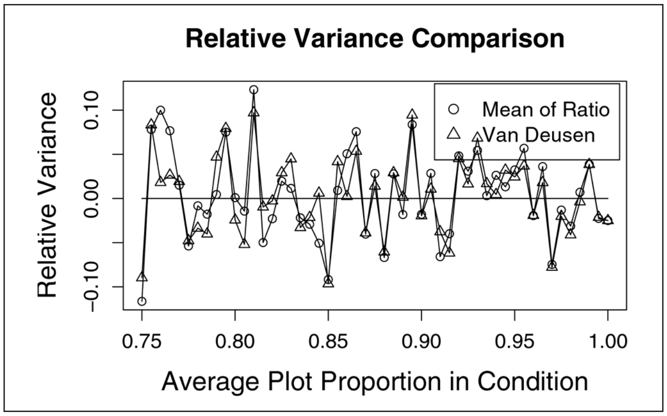
The simulation shows that the model that best describes the data produces the most accurate variance estimator, which is v_2 in equation (5b). The model that led to v_1 in equation (3b) is based on the notion that each plot measurement (y_i) has the same expected value regardless of the plot proportion (a_i) in the condition. Therefore, that this variance estimator performs poorly when plots have widely differing proportions within the condition is not surprising.

Other models might represent the data well that have not been considered here; therefore, no guarantee exists that equation (5b) is uniformly better than any alternative. For example, consider equation (4) with different variance assumptions on the error term, $e2_i$. Suppose the variance of the error was assumed to be a function of the squared proportion, i.e. $E(e2_i, e2'_i) = a_i^2 \sigma^2$. This would lead to the following estimate of the mean:

$$\bar{Y}_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{a_i} \quad (6a)$$

which is different from the mean estimator produced by the other models. A potential problem with this "mean of ratios" estimator is that plots with small a_i values are given large weights. Intuitively, this seems like a bad idea because these plots contain less information about the condition than a plot that is fully in the condition.

Figure 3.—*Simulated comparison of two variance estimators defined by equations (5b) and (6b). The relative variance is as defined for figure 2.*



The variance estimator for the ratio of means model is the following:

$$v(\bar{Y}_2) = \frac{\sum_{i=1}^{n} \dot{y}_i^2 - n\bar{Y}_2^2}{n(n-1)} \quad (6b)$$

where $\dot{y}_i = y_i / a_i$. This model was evaluated in the simulation described above, and it compared favorably (fig. 3) with equation (5b). The simulation included a_i values down to 0.5, so that no chance of dividing y_i by a miniscule a_i value was possible. Because equation (6b) doesn't perform any better than (5b) in this controlled situation, this equation is not recommended.

Variance Estimator Is Unbiased

The purpose of this section is to prove that the variance estimator for EST2 is unbiased. Reconsider the assumption that the variance of the error term in equation (4) is proportional to $a_i\sigma^2$. Suppose a circular plot contains the variable of interest, y. If the coverage of the plot is absolutely homogeneous, the amount of y, say f(y), on a proportion of the plot, would be f(y)=ay. This would lead to Var(f(y))=a^2Var(y). Recall that $0 \le a \le 1$ and therefore $a^2 \le a$. In fact, the y variable will not be exactly homogeneous across the plot; therefore, a^2Var(y) would understate the variance of f(y). A better approximation is likely to be Var(f(y))=aVar(y). This is the justification for the assumption used in equation (4).

Intuitively, adjusting the degrees of freedom downward when using partial plots makes sense, and thus the denominator of variance estimator (5c) uses $\sum a_i$ rather than n, the number of plots. Clearly, $\sum a_i \le n$, with equality occurring when all plots are fully in the condition. Adjusting the degrees of freedom downward for partial plots results in an unbiased estimate of the mapped plot variance. Consider the variance estimator for mapped plots with known population mean,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{\sum_{i=1}^{n} a_i} \quad (7)$$

The numerator of equation (7) could be written as $\sum e_i^2$ and, by the variance assumptions for equation (4), this has expected value of $\sigma^2 \sum a_i$. Therefore, equation (7) is an unbiased estimator. Equation (5c) is also justified by this result because it has 1 degree of freedom subtracted from the denominator to account for the estimated population mean parameter.

In general, the sample size for mapped plots should be adjusted to account for only using a proportion of the plot, a_i. This makes sense intuitively and theoretically as shown for equation (5c). A plot that is only 50 percent in the condition should count half as much as a plot that is fully in the condition. This is the justification for dividing by $\sum a_i$ rather than n in equation (5b).

Conclusions

Mapped plots are being installed by FIA as part of the annual forest inventory system. Two estimators for means and variance that have been proposed elsewhere were re-derived and compared. It turned out that both estimators for the mean were identical, but the variance estimators were quite different. A simulation showed that the variance estimator from the literature (Van Deusen 2004) performed significantly better than the alternative estimator.

A mean of ratios estimator was also discussed. The mean of ratios variance estimator performed about as well as the estimator from the literature (Van Deusen 2004). The mean of ratios estimator, however, could perform badly if small slivers of plots are included in the analysis, and, therefore, is not a recommended estimator for this particular use.

Literature Cited

Birdsey, R.A. 1995. A brief history of the "straddler plot" debates. Forest Science Monograph. 31: 7–11.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. New York: John Wiley and Sons.

Hahn, J.T.; MacClean, C.D.; Arner, S.L.; Bechtold, W.A. 1995. Procedures to handle inventory cluster plots that straddle two or more conditions. Forest Science Monograph. 31: 12–25.

Thompson, S.K. 2002. Sampling. New York: John Wiley and Sons.

Van Deusen, P.C. 2004. Forest inventory estimation with mapped plots. Canadian Journal of Forest Research. 34: 493–497.

Additional Readings

Bechtold, W.A.; Patterson, P.L., eds. 2004. The enhanced Forest Inventory and Analysis program—National sampling design and estimation procedures. http://srsfia2.fs.fed.us/publicweb/statistics_band/statband_documentation_rev14.1.doc.

Scott, C.T.; Bechtold, W.A. 1995. Techniques and computations for mapping plot clusters that straddle stand boundaries. Forest Science Monograph. 31: 46–61.