

Grid-Based Sampling Designs and Area Estimation

Joseph M. McCollum¹

Abstract.—The author discusses some area and variance estimation methods that have been used by personnel of the U.S. Department of Agriculture Forest Service Southern Research Station and its predecessors. The author also presents the methods of Horvitz and Thompson (1952), especially as they have been popularized by Stevens (1997), and shows how they could be used to produce estimates of variance on the fly from plots with static expansion factors. The author also extends the ideas of Horvitz and Thompson to the Forest Health Monitoring ozone grid.

Introduction

Bayesian analysts speak of “prior” and “posterior” distributions. The prior distribution is the initial estimate, and the posterior distribution is the corrected estimate. In the Forest Inventory and Analysis (FIA), the prior estimate is called phase 1, and is based solely on the dot count (if photointerpreters are estimating land cover) or pixel count (if remote sensing data is used). The posterior estimate is called phase 2, and is based on phase 1 but corrected for field calls.

An example that demonstrates area estimation procedures is given in the supplement (FIA documentation 2005) to Bechtold and Patterson (2005).

The phase 1 data is shown in table 1.

Thus, the phase 1 estimate is an $H \times 1$ vector $\mathbf{n}'$, where H is the number of strata, and its proportions in the $H \times 1$ vector $\mathbf{w}$. The field sites that are visited are held in an entirely different $H \times 1$ vector $\mathbf{n}$; in the context of this example, it is left to the field crew to decide how much of a plot is out of the population. The expansion factor—how much area a plot represents—is determined by A_T (114,000 acres for this hypothetical county), times n'_h divided by (n'_h) , where h is the appropriate subscript of each vector.

The phase 2 data is shown in table 2.

Thus, the plot count is held in $\mathbf{N}$, while the row-wise proportions are held in $\mathbf{W}$. For example, the number of plots that were photointerpreted to be forest and verified to be forest by the field crew is 7.3. Given that the photointerpreter called a dot forest, we find that there was an 83.9 percent chance that the field crew agreed. Given that the photointerpreter called a dot nonforest, we find that there was a 13.2 percent chance that the field crew found it to be forested. Given that the photointerpreter called a dot census water, we find that

Table 1.—Phase 1 data.

Phase 1 strata	Photo dots/pixel count	Stratum weight	Plot count, excluding out of population	Phase 1 expansion factor $A_T n'_h / (n'_h)$
1 (A) Forest land	$n'_1 = 75$	$w_1 = 0.5515$	$n_1 = 8.7$	7226.1663
2 (B) Nonforest land	$n'_2 = 44$	$w_2 = 0.3235$	$n_2 = 6.8$	5423.8754
3 (C) Census water	$n'_3 = 17$	$w_3 = 0.125$	$n_3 = 2$	7125
	$n' = 136$		$n = 17.5$	

¹ Information Technology Specialist, U.S. Department of Agriculture, Forest Service, Southern Research Station, 4700 Old Kingston Pike, Knoxville, TN 37919. E-mail: jmcollum@fs.fed.us.

Table 2.—*Confusion matrix.*

	Forest	Nonforest	Census water	Total
Forest	$n_{1,1} = 7.3$ $w_{1,1} = 0.839$	$n_{1,2} = 0.4$ $w_{1,2} = 0.046$	$n_{1,3} = 1$ $w_{1,3} = 0.115$	$n_{1,\bullet} = 8.7$
Photointerpretation nonforest	$n_{2,1} = 0.9$ $w_{2,1} = 0.132$	$n_{2,2} = 5.567$ $w_{2,2} = 0.819$	$n_{2,3} = 0.333$ $w_{2,3} = 0.049$	$n_{2,\bullet} = 6.8$
Call census	$n_{3,1} = 0.7$	$n_{3,2} = 0.5$	$n_{3,3} = 0.8$	$n_{3,\bullet} = 2$
Water	$w_{3,1} = 0.35$	$w_{3,2} = 0.25$	$w_{3,3} = 0.4$	
Total	$n_{\bullet,1} = 8.9$	$n_{\bullet,2} = 6.467$	$n_{\bullet,3} = 2.133$	$n_{\bullet,\bullet} = 17.5$

there was a 35 percent chance that the field crew found it to be forested. Thus, the phase 2 estimate is merely $\mathbf{p} = \mathbf{W}^T \mathbf{w}$, where matrix transposition is indicated by a superscript T, and one need only multiply by A_r to get $\mathbf{a}$, the land area estimate.

Doing this calculation, we obtain $\mathbf{p} = [0.5493 \ 0.3215 \ 0.1292]^T$, which in acres for this hypothetical county would be $\mathbf{a} = [62,620 \ 36,648 \ 14,732]^T$. Now, suppose we check the Census Bureau's gazetteer for this hypothetical county and it tells us that there are 59,569,965 m² (14,720 acres) of water in the county, representing 0.1291 of the county's total area. Alas, a contradiction is raised: the estimate is ever so slightly different from its known value.

The Statistics Band's proposed resolution to this contradiction may be found in the statement, "If census water is known (i.e., subtracted from A_r), condition classes in census water would be treated as out of the population, the same as plots that straddle national boundaries" (FIA documentation 2005). This statement seems to imply that merely dropping the census water column from $\mathbf{N}$ or $\mathbf{W}$ is enough, without any adjustment necessary to w_3 .

If we drop the census water column from $\mathbf{N}$ and then recompute $\mathbf{W}$, we get $\mathbf{a} = [63,615 \ 35,665 \ 14,720]^T$, almost a 1,000-acre increase in forest and nearly the same decrease in nonforest. Census water, however, should be filtered out of the $\mathbf{w}$ vector as well, although w_3 should not be reduced all the way to zero.

Thus, it is appropriate to ask, "What other prior distribution $\mathbf{v}$, most consistent with the existing prior distribution $\mathbf{w}$, when

multiplied into the confusion matrix ($\mathbf{W}^T$), gives a posterior with the known amount of census water?"

First, the total weight (sum of the vector's components) should be 1. Second, the weighted sum of census water should equal the amount of census water in the gazetteer. Third, the weights for the unknown components should be proportional to their phase 1 estimates. The following are the formal equations:

$$v_1 + v_2 + v_3 = 1 \quad (1)$$

$$w_{1,3} v_1 + w_{2,3} v_2 + w_{3,3} v_3 = p_3 = 0.1291 \quad (2)$$

$$v_1 - \frac{w_1}{w_2} v_2 = 0 \quad (3)$$

where:

w_{ij} = the weights from the confusion matrix $\mathbf{W}$, and

w_i = the weights from the phase 1 estimate $\mathbf{w}$.

Solve this system and get $\mathbf{v} = [0.55169 \ 0.32366 \ 0.12465]^T$, and now $\mathbf{p} = \mathbf{W}^T \mathbf{v} = [0.5494 \ 0.3215 \ 0.1291]^T$, while $\mathbf{a} = [62,629 \ 36,651 \ 14,720]^T$. If $\mathbf{w}$ differs greatly from $\mathbf{v}$, there are at least three possibilities. First, field crews may have misidentified census water as noncensus water, or vice versa. Second, there might be bias in the phase 1 estimate. For instance, digital photography may not cover the population. Producers of digital orthophotos may have deliberately excluded vast areas of territorial sea. Third, the estimate could be biased if the centers of the phase 1 cells are different from the remainder of the phase 1 cells.

Expansion Factors

Once **a** is estimated, these acres must be apportioned to plots by expansion factors. In work done by personnel of predecessors to the Southern Research Station, plots were not stratified according to photointerpretation. If anything, they were stratified by ownership, but ownership was determined by the field crews. Typically, expansion factors were calculated by dividing the estimate of forested acres by the number of forested plots. Return to the original solution, where **a** = [62,620 36,648 14,732]^T. Now divide 62,620 by the number of forested plots (as determined by the field crew), 8.9, to get an expansion factor of 7,036 acres per plot. A plot such as #14 in Supplement 5, or any such plot, that is 60 percent forested would have an expansion factor of 7,036 x 0.60 = 4,221 acres for this particular condition.

Meanwhile, there are 36,648 acres of nonforested land, divided by 6.467 nonforested plots (as determined by the field crew), and this yields an expansion factor of 5,667 acres per plot. A plot such as #14 in Supplement 5 that is 40 percent forested, or any such plot, would have an expansion factor of 5,667 x 0.40 = 2,267 acres.

With the two conditions taken together, plot #14 is apparently 4,221 acres / (4,221 acres + 2,267 acres) = 65 percent forested. This percentage is called the *adjusted condition proportion*.

Horvitz-Thompson

Horvitz and Thompson (1952) produced a fairly elegant method that does not require use of an adjusted condition proportion. Their method of estimating area is equivalent to that of the Statistics Band. Plot #14 in Supplement 5 was photointerpreted as forest. The field crew found the plot to be 60 percent forested. Thus, the expansion factor for that condition is 60 percent times the phase 1 acres (7,226, shown in table 1), or 4,336 acres. The expansion factor for the nonforested condition is 40 percent, or 2,890 acres.

Meanwhile, plot #3 was photointerpreted to be nonforest. The field crew found it to be 90 percent forested and 10 percent nonforested land, so the condition expansion factors are 4,882 and 542 acres, respectively. Again, there is no need for an adjusted condition proportion. The price for this simplicity is that plots in different strata now carry a different number of acres.

The total number of forested acres is the same in both methods, but the number of acres assigned to any particular plot is likely to differ from method to method. Thus, the estimate of area will be unaffected, but because acreage expansion factors are allotted differently, the estimates of volume, biomass, and total basal area, for instance, will be different.

Estimated Standard Errors

If there are only two strata of unknown proportions, the following equation for the variance of forested area, $s^2(p_1)$, may be used. In McCollum (2003), it was derived from Goodman (1960, 1962), although this equation was used before then. It may also be derived from Schumacher and Chapman (1954).

$$s^2(p_1) \approx \frac{w_1 w_2}{n_{..}} (w_{1,1} - w_{1,2})^2 + \frac{w_1^2 w_{1,1} w_{2,1}}{n_{1.}} + \frac{w_2^2 w_{1,2} w_{2,2}}{n_{2.}} \quad (4)$$

The Statistics Band offers two different formulae for variance estimation. One is for use with stratified random sampling:

$$v(\hat{A}_d) = \frac{A_T^2}{n} \left[\sum_h^H W_h n_h v(\bar{P}_{hd}) + \sum_h^H (1 - W_h) \frac{n_h}{n} v(\bar{P}_{hd}) \right] \quad (5)$$

where:

A_T = the total area of the area of interest (e.g., a county).

n = the total number of plots.

h = an index variable indicating the stratum.

H = the total number of strata.

W_h = the weight of stratum h (defined in this paper as w_h).

n_h = the number of plots in stratum h .

$\bar{P}_{hd}$ = the mean of the plot proportions in the domain of interest d assigned to stratum h .

The other formula was for use with double sampling for stratification:

$$v(\hat{A}_d) = A_T^2 \left\{ \sum_h^H \left(\frac{n'_h - 1}{n' - 1} \right) \frac{n'_h}{n'} v(\bar{P}_{hd}) + \frac{1}{n' - 1} \sum_h^H \frac{n'_h}{n} (\bar{P}_{hd} - \bar{P}_d)^2 \right\} \quad (6)$$

where:

n' = the total number of pixels.

n'_h = the number of pixels in stratum h .

$\bar{P}_d$ = the estimated proportion of the population in the domain of interest d .

All other symbols are as defined in equation (5).

The Statistics Band points out that these formulas are difficult, if not impossible, to implement with static expansion factors and arbitrary subsets of plots.

Inclusion Probabilities

Horvitz and Thompson (1952) as well as Yates and Grundy (1953) have developed a couple of estimators that could handle the problem of calculating standard errors on a subset of plots with static expansion factors. It would be necessary to attach the inclusion probability and the joint inclusion probability for a plot of each stratum to the plot record.

The Horvitz-Thompson estimator for variance is:

$$v(\hat{Y}_{HT}) = \sum_{i=1}^n \frac{(1 - \pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j \quad (7)$$

The Yates-Grundy estimator is:

$$v(\hat{Y}_{YG}) = \sum_{i=1}^n \sum_{j>i}^n \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (8)$$

where:

π_i = the inclusion probability for plot i .

π_{ij} = the joint inclusion probability for plots i and j .

y_i = the measured value on plot k (expanded acres, volume, biomass, etc.).

Cochran (1977) observes that:

$$\sum_{i=1}^{n'} \pi_i = n \quad (9)$$

and

$$\sum_{j \neq i}^{n'} \pi_{ij} = (n - 1) \pi_i \quad (10)$$

The fact that the inclusion probabilities and the joint inclusion probabilities do not sum to 1 is difficult to grasp at first, but in the example, if there are n' pixels or neighborhoods of dots in the population, and n of them are sampled, the inclusion probability will on average be equal to n / n' .

To estimate inclusion and joint inclusion probabilities, some assumptions have to be made. It is simplest to treat the plot list as a random selection from the list of pixels. Thus, a plot is understood to be no larger than a pixel, and pixels are assumed to be independent. In reality, adjacent Landsat Thematic Mapper pixels are not independent; a subplot is about one-fifth the size of such a pixel.

Under these assumptions, inclusion probability for plot i in stratum h is:

$$\pi_i = \frac{q_{h(i)}}{n'_{h(i)}} \quad (11)$$

where $h(i)$ is a function assigning plot i to stratum h .

Joint inclusion probability for plot i in stratum h and plot j in a different stratum, k , is:

$$\pi_{i,j} = \frac{q_{h(i)}}{n'_{h(i)}} \cdot \frac{(q_{h(j)} - r)}{(n'_{h(j)} - t)} \quad (12)$$

For i and j in different strata, r will typically equal 0 (because no plot has been removed from the second stratum), and for i and j in the same stratum, r will typically equal 1. The simplest assumption for a model without replacement is that $t = 0$ for plots i and j in different strata, and $t = 1$ for plots i and j in the same stratum. This assumption produces the largest estimate of variance for Horvitz-Thompson and Yates-Grundy. This estimate can be improved on, however. There are 27 phase 1

dots per phase 2 cell, and it is known to which strata the dots belong. Thus, t can equal the number of phase 1 dots belonging to the same stratum as plot j .

Ozone Grid

Another use of the Horvitz-Thompson estimator would involve analysis of the current ozone data. The Horvitz-Thompson estimator is:

$$\hat{y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} \quad (13)$$

For the 2002 field season, ozone symptoms were collected on an entirely different grid (Smith *et al.* 2001). The four strata were as follows:

Stratum 0: 1 ozone plot per 5,862,400 acres (one plot per 256/7 historic phase 3 cells).

Stratum 1: 1 ozone plot per 1,465,600 acres (one plot per 64/7 historic P3 cells).

Stratum 2: 1 ozone plot per 1,139,911 acres (one plot per 64/9 historic P3 cells).

Stratum 3: 1 ozone plot per 641,200 acres (one plot per 4 historic P3 cells).

Expansions and contractions of the grid are easy to do if the factors are of the form $T = h^2 + hk + k^2$; frequently used factors are $T = 3$ ($h = 1, k = 1$), 4 ($h = 2, k = 0$), and 7 ($h = 2, k = 1$).

There are 27 phase 2 plots per historic P3 cell. The current ratio is 16.0:1 in some States and 16.2:1 in others. (McCollum and Cochran 2005).

Stratum 0 included all areas that were < 7.5 percent forest, plus all areas that were > 90 percent pinyon-juniper, a species not sensitive to ozone. Stratum 1 included all areas of low-ozone risk. Stratum 2 included areas of moderate ozone risk. Stratum 3 included all areas of high-ozone risk.

With such a sample design, it is inappropriate to report results based on a raw plot count. A weighted average is far more appropriate. Stevens (1997) called a sample design similar to this one a multidensity, randomized-tesselation, stratified design. Inclusion probabilities could be calculated by dividing the number of ozone plots in each risk stratum by the number of phase 1 photointerpretation dots that fall in that stratum. If tabulating the number of phase 1 dots in each ozone risk stratum is impractical, then an alternative could be tabulating the number of phase 2 plots in each ozone risk stratum. If there is no difference between strata in terms of actual ozone risk and it is proper to use raw plot count for the ozone grid, then the sample design does not capture areas of ozone risk.

Conclusions

First, a retrieval system ought to be able to incorporate Horvitz-Thompson or Yates-Grundy variance estimators easily, although it is not clear that an ordinary user could construct either estimator. Other variance estimators may also be constructed on the fly.

Second, the author recommends against abandoning expansion factors, and in favor of keeping static expansion factors. Separate expansion factors will be required for inventory and remeasurement, and the old inventory plus the remeasurement will not equal the new inventory. Expansion factors should be based on the entire cycle. To get an unbiased estimate from one panel or one subcycle of data, the author points out that remeasurement data should be available, and it could be noted what plots have been dropped since the last cycle.

Third, the author recommends that census water be enumerated in the manner set forth in this paper.

Last, the author recommends that the ozone data be analyzed by stratum.

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. http://srsfia2.fs.fed.us/publicweb/statistics_band/stat_documents.htm. (21 December 2005).
- Cochran, W.G. 1977. Sampling techniques. 3rd ed. New York: John Wiley. 428 p.
- FIA documentation. 2006. Forest Inventory and Analysis (FIA) statistics band documentation. http://srsfia2.fs.fed.us/publicweb/statistics_band/stat_documents.htm. (21 December 2005).
- Goodman, L.A. 1960. On the exact variance of products. *Journal of the American Statistical Association*. 55: 708-713.
- Goodman, L.A. 1962. The variance of the product of k random variables. *Journal of the American Statistical Association*. 59: 54-60.
- Horvitz, D.G.; Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 47: 663-685.
- McCollum, J.M. 2003. Standard errors in forest area. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; Moser, J.W., eds. Proceedings, third annual forest inventory and analysis symposium. Gen. Tech. Rep. NC-230. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station. 208 p.
- McCollum, J.M.; Cochran, J.K. 2005. Forest inventory and analysis and forest health monitoring: piecing the quilt. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; McWilliams, W.H., eds. Proceedings, the fifth annual Forest Inventory and Analysis symposium; 2003 November 18–20; New Orleans, LA. Gen. Tech. Rep. WO-69. Washington, DC: U.S. Department of Agriculture, Forest Service: 175-182.
- Schumacher, F.X.; Chapman, R.A. 1954. Sampling methods in forestry and range management. Durham, NC: Duke University School of Forestry. 222 p.
- Smith, W.D.; Coulston, J.; Smith, G. 2001. Development of the ozone indicator sampling grid. Unpublished data. Author can be reached at U.S. Department of Agriculture, Forest Service, Southern Research Station, Forestry Sciences Laboratory, P.O. Box 12254, Research Triangle Park, NC 27709.
- Stevens, D.L. 1997. Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics*. 8(3): 167-195.
- Yates, F.; Grundy, P.M. 1953. Selection without replacement from within strata with probability proportion to size. *Journal of the Royal Statistical Society. Series B*. (1): 253-261.