REGRESSION WEIGHTING FOR MULTIPHASE FOREST SERVICE SAMPLES

F. Jay Breidt and Wayne A. Fuller, Iowa State University
F. Jay Breidt, 221 Snedecor Hall, Ames, IA 50011

KEY WORDS: Regression estimation, resource inventory, jackknife, multiple phase sample.

Abstract: Some natural resource inventories conducted by the United States Forest Service in Alaska consist of multiple phases of sampling, with observations on the sampling units obtained from satellite imagery, high and low altitude photography and ground visits, for example. For such samples, we suggest estimators constructed in two steps. First, standard regression estimation on data collected in all but the last phase of sampling provides initial estimates. Second, a regression weight generation procedure is applied to data collected in the final phase of sampling. Estimates from the first step are used as control totals. The generated weights lead to simple, internally consistent estimators. The methods discussed are applied to three-phase data from the Tanana region of eastern Alaska. Jackknife variance estimation is employed to compare the efficiency of the full multiphase estimator to the simplified estimator.

1 Introduction

We outline two estimation procedures for the multiphase samples conducted by the Forest Service in Alaska and apply these procedures to a resource inventory of the Tanana river basin in eastern Alaska. Though the Tanana survey was conducted as a four-phase sample, with imagery from the Landsat multispectral scanner satellite collected in the first phase, we consider only three phases. Thomas (1978) considers a three-phase sample in a similar remote sensing application, adapting the results of Tikkiwal (1967) for estimation.

Consider a finite universe of N elements, denoted U, where the elements of U are 20-acre plots of land. A Phase I probability sample of n_I plots, denoted I, is drawn from U and a high altitude photograph (HAP) is collected for each plot. The information on the high altitude photographs is the percentages of the 20-acre plot classified as falling in different cover categories, such as needle leaf forest, permanent ice and snow, water, etc. Denote this vector of information by

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \quad i = 1, \ldots, n_I,$$

where

$$x_{ij} = \text{percentage of high-altitude photograph } i \text{ in cover category } j.$$

A Phase II subsample of n_{II} plots, denoted II, is drawn from I and a low altitude photograph (LAP) is collected for each Phase II plot. Cover category percentages are also collected from the low altitude photographs, yielding the information vector

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iq}) \quad i = 1, \ldots, n_{II},$$

where

$$y_{ij} = \text{percentage of low-altitude photograph } i \text{ in cover category } j.$$

Finally, a Phase III subsample of n_{III} plots, denoted III, is drawn from II and a ground visit (GRD) is conducted for each Phase III plot. During ground observation, a number of points within the plot are visited and the cover category is recorded for each. There are 19 points in each GRD plot in the Tanana region. The data are

$$\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ir}) \quad i = 1, \ldots, n_{III},$$

where

$$z_{ij} = \text{percentage of 19 ground points in plot } i \text{ with cover category } j.$$

Note that p, q and r generally differ, since different features can be resolved in different phases. In the estimation for Tanana, we used $p = 7$ Phase I controls and $q = 7$ Phase II controls. In both phases, these controls were percentages in cover categories called needle leaf forest, shrub, mixed forest, broadleaf forest, rock and alluvial, permanent ice and snow, and water. Although the same names are used, cover categories in different phases are not directly comparable—what appears to be ice and snow in HAP may not appear to be ice and snow in LAP and may not be ice and snow in a GRD visit. Early phases of sampling do, however, provide useful auxiliary information.

Roughly, for the samples we consider, the Phase I plots are located on a 10-km square grid, the Phase II plots on a 20-km subgrid and the Phase III plots on a further 40-km subgrid (Schreuder, Gregoire and Wood 1993, pp. 189–190). Thus, there are about four times as many Phase I plots as Phase II plots and four times as many Phase II plots as Phase III plots. For example, the Tanana river basin, an area of about 34.56 million acres, contains $n_I = 1290$ Phase I plots, $n_{II} = 323$ Phase II plots and $n_{III} = 86$ Phase III plots. (For simplicity, all plots with missing data were removed from the analysis). The design for Tanana is shown in Figure 1.

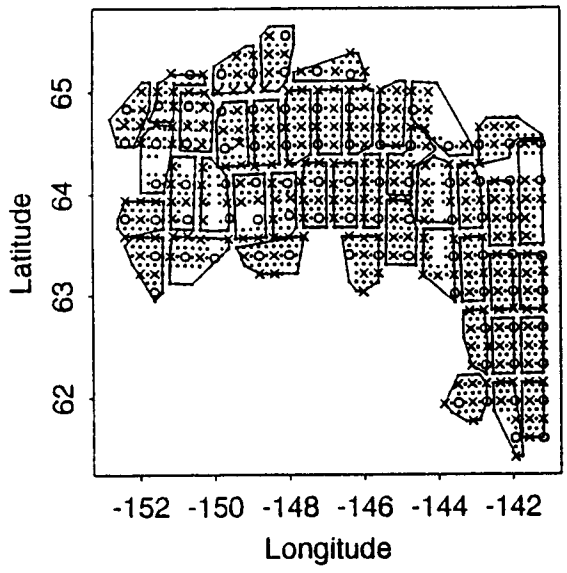

Figure 1: *Three-phase sampling design (· = HAP only; × = LAP and HAP; ∘ = GRD, LAP and HAP) for Forest Service resource inventories in the Tanana river basin region, Alaska. Solid lines show strata constructed for the purpose of variance estimation.*

2 Estimation

A full three-phase estimator may be derived under simple random sampling as the maximum likelihood estimator (MLE) of the mean vector in a multivariate normal framework. More general sampling designs could be treated by following the unequal probability development in Särndal and Swensson (1987).

Because of the special nesting structure of the $\mathbf{x}_i$, $\mathbf{y}_i$ and $\mathbf{z}_i$ data, MLE's of the multivariate normal parameters (including the covariance matrices Σ_{xx}, Σ_{xy}, Σ_{xz}, Σ_{yy}, Σ_{yz} and Σ_{zz}) may be derived using a method described by Anderson (1957). MLE's of the mean vectors for $\mathbf{x}_i$, $\mathbf{y}_i$ and $\mathbf{z}_i$ are given by

$$\begin{aligned}
\hat{\mu}_x &= \bar{\mathbf{x}}_I, \\
\hat{\mu}_y &= \bar{\mathbf{y}}_{II} + (\hat{\mu}_x - \bar{\mathbf{x}}_{II})\hat{\beta}_{y\cdot x}, \\
\hat{\mu}_z &= \bar{\mathbf{z}}_{III} + \left[\, \hat{\mu}_x - \bar{\mathbf{x}}_{III} \quad \hat{\mu}_y - \bar{\mathbf{y}}_{III} \,\right] \hat{\beta}_{z\cdot xy},
\end{aligned}$$

where

$$\bar{\mathbf{x}}_I = \frac{\sum_I \mathbf{x}_i}{n_I},\ \bar{\mathbf{x}}_{II} = \frac{\sum_{II} \mathbf{x}_i}{n_{II}},\ \bar{\mathbf{x}}_{III} = \frac{\sum_{III} \mathbf{x}_i}{n_{III}},$$
$$\bar{\mathbf{y}}_{II} = \frac{\sum_{II} \mathbf{y}_i}{n_{II}},\ \bar{\mathbf{y}}_{III} = \frac{\sum_{III} \mathbf{y}_i}{n_{III}},$$
$$\bar{\mathbf{z}}_{III} = \frac{\sum_{III} \mathbf{z}_i}{n_{III}},$$

$\hat{\beta}_{y\cdot x}$ is the $p \times q$ matrix of estimated coefficients for the regression of $\mathbf{y}_i$ on $\mathbf{x}_i$ and $\hat{\beta}_{z\cdot xy}$ is the $(p+q)\times r$ matrix of estimated coefficients for the regression of $\mathbf{z}_i$ on $(\mathbf{x}_i, \mathbf{y}_i)$. (Intercepts are omitted in these coefficient matrices.) Note that $\hat{\mu}_y$ is the usual two-phase or double sample regression estimator under simple random sampling at each phase (e.g., Cochran 1977, p. 339).

Computationally, an attractive alternative to $\hat{\mu}_z$ is the three-phase estimator computed by first regressing $\mathbf{y}_i$ on $\mathbf{x}_i$ in Phase II and obtaining n_{II} vectors of deviations,

$$\begin{aligned}
\mathbf{d}_i &= \mathbf{y}_i - \hat{\nu}_y - \mathbf{x}_i\hat{\beta}_{y\cdot x} \\
&= (d_{i1}, d_{i2}, \ldots, d_{iq}),
\end{aligned}$$

where $\hat{\nu}_y$ is the intercept in the regression of $\mathbf{y}_i$ on $\mathbf{x}_i$.

Using the same style of argument as for the MLE, a three-phase estimator is given by

$$\tilde{\mu}_z = \bar{\mathbf{z}}_{III} + \begin{bmatrix} (\hat{\mu}_x - \bar{\mathbf{x}}_{III})' \\ (\bar{\mathbf{d}}_{II} - \bar{\mathbf{d}}_{III})' \end{bmatrix}' \hat{\beta}_{z \cdot xd},$$

where $\bar{\mathbf{d}}_{II}$ and $\bar{\mathbf{d}}_{III}$ are the means of $\mathbf{d}_i$ over II and III and $\hat{\beta}_{z \cdot xd}$ is the $(p+q) \times r$ matrix of estimated coefficients for the regression of $\mathbf{z}_i$ on $(\mathbf{x}_i, \mathbf{d}_i)$. (Intercepts are omitted in this coefficient matrix.)

It is convenient to construct a regression weight for each ground plot, since then estimators built upon ground plot characteristics can be constructed with ease. The regression weights are

$$\tilde{w}_i = \frac{1}{n_{III}} + \begin{bmatrix} 1-1 \\ (\hat{\mu}_x - \bar{\mathbf{x}}_{III})' \\ (\bar{\mathbf{d}}_{II} - \bar{\mathbf{d}}_{III})' \end{bmatrix}' (F'F)^{-1} \begin{bmatrix} 1 \\ \mathbf{x}_i' \\ \mathbf{d}_i' \end{bmatrix},$$

where F is the $n_{III} \times (1+p+q)$ matrix

$$F = \begin{bmatrix} 1 & \mathbf{x}_i & \mathbf{d}_i \end{bmatrix}_{i=1}^{n_{III}}.$$

These weights have the property

$$\sum_{III} \tilde{w}_i \begin{bmatrix} 1 & \mathbf{x}_i & \mathbf{d}_i \end{bmatrix} = \begin{bmatrix} 1 & \hat{\mu}_x & \bar{\mathbf{d}}_{II} \end{bmatrix};$$

i.e., the weights sum to one, the weighted $\mathbf{x}_i$ data from Phase III sum to the $\mathbf{x}_i$ control totals from Phase I and the weighted $\mathbf{d}_i$ data from Phase III sum to the $\mathbf{d}_i$ control totals from Phase II.

A simplified regression estimator can be constructed in two steps. In the first step, compute the two-phase regression estimator for the population mean of the $\mathbf{y}_i$'s, $\hat{\mu}_y$. In the second step, use $\hat{\mu}_y$ as a vector of control totals in a regression of $\mathbf{z}_i$ on $\mathbf{y}_i$. The two-step estimator is

$$\mu_z^* = \bar{\mathbf{z}}_{III} + (\hat{\mu}_y - \bar{\mathbf{y}}_{III})\hat{\beta}_{z \cdot y},$$

where $\hat{\beta}_{z \cdot y}$ is the $q \times r$ matrix of estimated coefficients for the regression of $\mathbf{z}_i$ on $\mathbf{y}_i$. (Intercepts are omitted in this coefficient matrix.) Kiregyera (1984), following Chand (1975), suggested an estimator like μ_z^* using a single variable at each phase.

The regression weights are

$$w_i^* = \frac{1}{n_{III}} + \begin{bmatrix} 1-1 \\ (\hat{\mu}_y - \bar{\mathbf{y}}_{III})' \end{bmatrix}' (G'G)^{-1} \begin{bmatrix} 1 \\ \mathbf{y}_i' \end{bmatrix},$$

where G is the $n_{III} \times (1+q)$ matrix

$$G = \begin{bmatrix} 1 & \mathbf{y}_i \end{bmatrix}_{i=1}^{n_{III}}.$$

These weights have the property

$$\sum_{III} w_i^* \begin{bmatrix} 1 & \mathbf{y}_i \end{bmatrix} = \begin{bmatrix} 1 & \hat{\mu}_y \end{bmatrix};$$

i.e., the weights sum to one and the weighted Phase III $\mathbf{y}_i$ data sum to the control totals $\hat{\mu}_y$.

Both the three-phase and the two-step regression weights lead to internally consistent estimates for the means of Phase III measurements, in the sense that the estimate for a linear combination of parameters is given by the linear combination of the parameter estimates. The three-phase estimator, $\tilde{\mu}_z$, would be optimal in some sense if the regression coefficient matrices were known, but the large number of regression parameters to be estimated, $pq + (p+q)r$, reduces the efficiency of the estimator. The two-step estimator, μ_z^*, requires the estimation of only $pq + qr$ regression coefficients. A hypothesis of this work is that, largely due to parameter estimation variability, the two-step estimator can be competitive with the three-phase estimator.

3 Jackknife Variance Estimation

Since p and q were quite large relative to the number of Phase III observations, estimation error was expected to contribute significantly to the variability of the regression estimators. Jackknife variance estimation was used to compare the efficiencies of the estimators because

it can reflect the error due to estimating regression coefficients. A naive jackknife approach would be to treat this multiphase systematic sample as a simple random sample at each phase, but the presence of relatively strong stratification effects for some variables makes the naive jackknife standard errors unstable in the sense that they are very sensitive to the number of observations deleted for each pseudo-replicate. For this reason, we treated this multiphase systematic sample as a stratified cluster sample with 43 strata and two clusters per stratum.

Strata were constructed (using the "brush" function in the S-Plus software) so that each stratum contained two Phase III plots, an average of 7.5 Phase II plots and an average of 30 Phase I plots. Figure 1 shows the 43 strata. Clusters were formed by ordering the plots on latitude within strata and taking the northern and southern "halves" as clusters. Each cluster contained approximately 15 Phase I plots, approximately 3.8 Phase II plots and exactly one Phase III plot.

Let $\tilde{\mu}_{z,(hg)}$ and $\mu^*_{z,(hg)}$ denote the estimators of the same functional form as $\tilde{\mu}_z$ and μ^*_z, respectively, computed after deleting the gth cluster of the hth stratum ($g = 1, 2$; $h = 1, \ldots, 43$). Jackknife pseudovalues are then given by

$$\tilde{\mu}_{z,hi} = 44\tilde{\mu}_z - 43\tilde{\mu}_{z,(hg)}$$

and

$$\mu^*_{z,hi} = 44\mu^*_z - 43\mu^*_{z,(hg)}$$

(Wolter, 1985, §4.5). The jackknife estimators are

$$\tilde{\mu}_{z,JK} = \sum_{g=1}^{2}\sum_{h=1}^{43}\tilde{\mu}_{z,hi}/86$$

and

$$\mu^*_{z,JK} = \sum_{g=1}^{2}\sum_{h=1}^{43}\mu^*_{z,hi}/86,$$

and jackknife variance estimators are given by the diagonal elements of the $r \times r$ matrices

$$\tilde{v}_{JK} = \sum_{h=1}^{43}(\tilde{\mu}_{z,(h1)} - \tilde{\mu}_{z,(h2)})'(\tilde{\mu}_{z,(h1)} - \tilde{\mu}_{z,(h2)})$$

and

$$v^*_{JK} = \sum_{h=1}^{43}(\mu^*_{z,(h1)} - \mu^*_{z,(h2)})'(\mu^*_{z,(h1)} - \mu^*_{z,(h2)})$$

(Wolter, 1985, §4.5).

4 Comparison of the Regression Estimators

Table 1 shows estimated percentages and jackknife standard errors for both $\tilde{\mu}_z$ and μ^*_z. The estimated percentages agree closely for all of the cover categories. According to the jackknife standard errors, the simplified estimator μ^*_z performs as well or better than the full three-phase estimator, $\tilde{\mu}_z$, for most cover categories.

A notable exception is for census water, the category consisting of lakes at least 40 acres in area and streams at least 0.125 miles in width, which was the rarest of the categories considered. The efficiency of the two-step estimator of census water is greatly improved by the use of percentages of high and low altitude photographs in census water as control totals in Phase I and in Phase II. In this example, use of census water as a control at the two phases led to a near-singularity in the computation of the three-phase estimator. Thus, the two-step estimator allowed greater flexibility in the choice of regressors than did the three-phase estimator.

It is also worth remarking that Phase III plots with HAP data missing could easily be incorporated into the two-step estimator. This advantage over the three-phase estimator was not exploited here, though one Phase III plot did have missing HAP data.

Acknowledgements: This work was supported by the U.S. Department of Agriculture's Soil Conservation Service under SCS Cooperative Agreement No. 68–3A75–2–64. We thank the members of the Forest Inventory Analysis Unit, U.S. Forest Service, Anchorage, Alaska for their cooperation in this project; in particular, we thank former member James Labau and current members Frederic Larson and Gary Carroll.

References

Anderson, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200–203.

Chand, L. (1975). Some ratio-type estimators based on two or more auxiliary variables. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa USA.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

Kiregyera, B. (1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. *Metrika* **31**, 215–226.

Särndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* **55**, 279–294.

Schreuder, H.T., Gregoire, T.G. and Wood, G.B. (1993). *Sampling Methods for Multiresource Forest Inventory.* Wiley, New York.

Thomas, R.W. (1978). Multiphase crop acreage estimation incorporating Landsat data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 169–174.

Tikkiwal, B.D. (1967). Theory of multiphase sampling from a finite or an infinite population on successive occasions 1, 2. *Review of the International Statistical Institute* **35**, 247–263.

Wolter, K.M. (1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

Cover Category	Estimated Percentage	
	$\tilde{\mu}_z$ (JK s.e.)	μ_z^* (JK s.e.)
Needle leaf forest	32.482	32.389
	(2.782)	(2.430)
Shrub	19.714	19.642
	(2.431)	(2.480)
Woodlands	11.747	11.360
	(2.238)	(2.106)
Mixed forest	10.421	10.565
	(2.284)	(2.313)
Broadleaf forest	7.689	7.937
	(2.481)	(2.615)
Rock and alluvial	5.420	5.401
	(0.778)	(0.814)
Ice and snow	5.049	5.049
	(1.246)	(1.246)
Herbaceous (dry)	3.867	3.981
	(1.577)	(1.641)
Herbaceous (wet)	1.652	1.725
	(0.838)	(0.962)
Small water	1.252	1.190
	(0.499)	(0.530)
Census water	0.659	0.719
	(0.395)	(0.576)

Table 1: *Comparison of components of* $\tilde{\mu}_z$ *and* μ_z^* *and their jackknife standard errors (JK s.e.'s) for different cover categories.*