AREAL CONTROL USING GENERALIZED LEAST SQUARES AS AN ALTERNATIVE TO STRATIFICATION[1]

Raymond L. Czaplewski[2]

Abstract—Stratification for both variance reduction and areal control proliferates the number of strata, which causes small sample sizes in many strata. This might compromise statistical efficiency. Generalized least squares can, in principle, replace stratification for areal control.

INTRODUCTION

Stratification with remotely sensed forest map improves precision of FIA estimates. However, FIA also uses stratification to assure that area estimates equal “official” acres in each county, as published by the U.S. Census Bureau. I term this latter function “areal control.”

Cross-stratification by both remotely sensed strata and geopolitical boundaries proliferates the number of strata. Hence, sample sizes become small in many of these cross-classified strata. There is concern within FIA that these small sample sizes can degrade the statistical efficiency gained through stratification. This concern is heightened by the move to an annual FIA system, in which only 10 percent to 15 percent of the FIA field plots are remeasured each year.

I consider the use of stratification solely for variance reduction to avoid proliferation of strata. I present an alternative to stratification to areal control that constrains FIA estimates such that they agree with county acreages from the Census Bureau. I use a simple example of two strata (forest and nonforest) and two counties.

PROBLEM FORMULATION

Let the 4x1 vector **z** represent the true area of forest and nonforest in each county. Equation (1) denotes the vector estimate of these areas, including the 4x4 covariance matrix $\mathbf{V_z}$ for estimation errors.

$$\hat{\mathbf{z}}=\left[\begin{array}{l}\hat{z}_1=\begin{array}{l}\text{Estimate of forest in county A}\\ \text{using remotely sensed strata}\end{array}\\ \hline \hat{z}_2=\begin{array}{l}\text{Estimate of nonforest in county}\\ \text{using remotely sensed strata}\end{array}\\ \hline \hat{z}_3=\begin{array}{l}\text{Estimate of forest in county B}\\ \text{using remotely sensed strata}\end{array}\\ \hline \hat{z}_4=\begin{array}{l}\text{Estimate of forest in county B}\\ \text{using remotely sensed strata}\end{array}\end{array}\right] \quad (1)$$

where

$$\mathbf{z} = \hat{\mathbf{z}} + \mathbf{u_z}$$
$$E\left[\mathbf{u_z u_z'}\right] = \mathbf{V_z}$$

The estimates in equation (1) assume that remotely sensed data have already been used to separate Phase 2 plots into forest and nonforest strata, and the appropriate estimator is used to reduce variance through this stratification. However, the estimates in equation (1) are not stratified by Census Bureau county statistics for areal control. The following describes the alternative to stratification for areal control, in which the sample estimate is constrained so that summations of areal estimates for each county exactly equal the “official” acres in each county, as published by the U.S. Census Bureau.

Let the 2x1 vector **c** contain the “official” acres in each county. A sample estimate of **c** is available from a simple linear transformation of the vector estimate **z** from equation (1).

$$\hat{\mathbf{c}} = \left[\begin{array}{c|c|c|c}1&1&0&0\\ \hline 0&0&1&1\end{array}\right]\left[\begin{array}{c}\hat{z}_1\\ \hline \hat{z}_2\\ \hline \hat{z}_3\\ \hline \hat{z}_4\end{array}\right] = \mathbf{H}\hat{\mathbf{z}} \qquad (2)$$

where

$$\mathbf{c} = \hat{\mathbf{c}} + \mathbf{u}_c$$
$$E\left[\mathbf{u_c u_c'}\right] = \mathbf{HV_zH'}$$

In addition, the exact areas for each county are available from the Census Bureau.

$$\mathbf{c}=\left[\begin{array}{l}c_1=\{\text{Census Bureau area in county A}\}\\ \hline c_2=\{\text{Census Bureau area in county B}\}\end{array}\right] \qquad (3)$$

c in equation (3) is a vector of constants, not an estimate, because the county acreages from the Census Bureau are known without error. Our objective is to constrain the vector estimate **z** in equation (1) such that the vector estimate **c** of county acreages in equation (2) agrees exactly with the Census Bureau statistics in equation (3).

GENERALIZED LEAST SQUARES (GLS)

Let vector **B** represent the estimates of forest area that are constrained to agree with official statistics.

[1] Paper presented at the Second Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Project Leader, USDA Forest Service, Rocky Mountain Research Station, 2150 Centre Ave., Bldg. A, Suite 350, Fort Collins, CO 80526-1891.

$$\mathbf{B} = \begin{bmatrix} \beta_1 \\ \hline \beta_2 \\ \hline \beta_3 \\ \hline \beta_4 \end{bmatrix} = \begin{bmatrix} \beta_1 = \text{acres of forest cover in county A} \\ \hline \beta_2 = \text{acres of non-forest cover in county A} \\ \hline \beta_3 = \text{acres of forest cover in county B} \\ \hline \beta_4 = \text{acres of non-forest cover in county A} \end{bmatrix} \tag{4}$$

such that

$$\beta_1 + \beta_2 = \text{Exact Census Bureau acres in county A}$$
$$\beta_1 + \beta_2 = \text{Exact Census Bureau acres in county B}$$

Estimation of the **B** vector is the final goal.

The GLS solution estimates 4x1 vector **B** in Equation (4) from the 6x1 vector **y,** which is a concatenation of vector estimate **z** from equation (1), and a vector of independent ancillary estimates **k** (with covariance matrix $\mathbf{V}_k$), which will later contain the areal control statistics.

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{\mathbf{z}} \\ \hat{\mathbf{k}} \end{bmatrix} = \begin{bmatrix} \hat{z}_1 \\ \hline \hat{z}_2 \\ \hline \hat{z}_3 \\ \hline \hat{z}_4 \\ \hline \hat{k}_1 \\ \hline \hat{k}_2 \end{bmatrix} \tag{5}$$

Define the linear model:

$$\mathbf{y} = \mathbf{XB} + \mathbf{u}_y \tag{6}$$

where

$$\mathbf{X} = \left[\begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 \end{array}\right] = \begin{bmatrix} \mathbf{I} \\ \hline \mathbf{H} \end{bmatrix}$$

$$\mathbf{V}_y = E\left[\mathbf{u}_y \mathbf{u}_y'\right] = \left[\begin{array}{c|c} \mathbf{V}_z & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{V}_k \end{array}\right]$$

The zero off-diagonal sub-matrices within the covariance matrix $\mathbf{V}_y$ in equation (6) correspond to the presumed independence between vectors **z** and **k** in equation (5).

The GLS estimator for **B** is defined as:

$$\hat{\mathbf{B}} = \left[\mathbf{X}' \left(\hat{\mathbf{V}}_y\right)^{-1} \mathbf{X}\right]^{-1} \mathbf{X}' \left(\hat{\mathbf{V}}_y\right)^{-1} \hat{\mathbf{y}} \tag{7}$$

Let the matrices in equation (7) be partitioned as follows:

$$\hat{\mathbf{B}} = \left[\begin{pmatrix} \mathbf{I} \\ \hline \mathbf{H} \end{pmatrix}' \left(\begin{array}{c|c} \hat{\mathbf{V}}_z & 0 \\ \hline 0 & \hat{\mathbf{V}}_k \end{array}\right)^{-1} \begin{pmatrix} \mathbf{I} \\ \hline \mathbf{H} \end{pmatrix}\right]^{-1} \begin{pmatrix} \mathbf{I} \\ \hline \mathbf{H} \end{pmatrix}' \left(\begin{array}{c|c} \hat{\mathbf{V}}_z & 0 \\ \hline 0 & \hat{\mathbf{V}}_k \end{array}\right)^{-1} \begin{bmatrix} \hat{\mathbf{z}} \\ \hline \mathbf{k} \end{bmatrix} \tag{8}$$

Using matrix algebra for partitioned matrices, equation may be rewritten as:

$$\begin{aligned} \hat{\mathbf{B}} &= \left[\begin{pmatrix} \mathbf{I} & \mathbf{H}' \end{pmatrix} \left(\begin{array}{c|c} \hat{\mathbf{V}}_z^{-1} & 0 \\ \hline 0 & \hat{\mathbf{V}}_k^{-1} \end{array}\right) \begin{pmatrix} \mathbf{I} \\ \hline \mathbf{H} \end{pmatrix}\right]^{-1} \begin{pmatrix} \mathbf{I} & \mathbf{H}' \end{pmatrix} \left(\begin{array}{c|c} \hat{\mathbf{V}}_z^{-1} & 0 \\ \hline 0 & \hat{\mathbf{V}}_k^{-1} \end{array}\right) \begin{bmatrix} \hat{\mathbf{z}} \\ \hline \mathbf{k} \end{bmatrix} \\ &= \left[\begin{pmatrix} \mathbf{I} & \mathbf{H}' \end{pmatrix} \begin{pmatrix} \hat{\mathbf{V}}_z^{-1} \\ \hat{\mathbf{V}}_k^{-1} \mathbf{H} \end{pmatrix}\right]^{-1} \begin{pmatrix} \mathbf{I} & \mathbf{H}' \end{pmatrix} \begin{pmatrix} \hat{\mathbf{V}}_z^{-1} \hat{\mathbf{z}} \\ \hat{\mathbf{V}}_k^{-1} \mathbf{k} \end{pmatrix} \\ &= \left[\hat{\mathbf{V}}_z^{-1} + \mathbf{H}'\hat{\mathbf{V}}_k^{-1} \mathbf{H}\right]^{-1} \left[\hat{\mathbf{V}}_z^{-1} \hat{\mathbf{z}} + \mathbf{H}'\hat{\mathbf{V}}_k^{-1} \mathbf{k}\right] \end{aligned} \tag{9}$$

Maybeck (1979, pp. 234-235) shows that the result in equation (9) is a maximum likelihood estimate under appropriate assumptions. Maybeck (p. 214) uses the following matrix inversion lemma to rewrite the first term in equation (9) into a numerically superior form:

$$\begin{aligned} \left[\hat{\mathbf{V}}_z^{-1} + \mathbf{H}'\hat{\mathbf{V}}_k^{-1} \mathbf{H}\right]^{-1} &= \hat{\mathbf{V}}_z - \hat{\mathbf{V}}_z \mathbf{H}' \left[\mathbf{H}\hat{\mathbf{V}}_z \mathbf{H}' + \hat{\mathbf{V}}_k\right]^{-1} \mathbf{H}\hat{\mathbf{V}}_z \\ &= \hat{\mathbf{V}}_z - \mathbf{GH}\hat{\mathbf{V}}_z \\ \text{where } \mathbf{G} &= \hat{\mathbf{V}}_z \mathbf{H}' \left[\mathbf{H}\hat{\mathbf{V}}_z \mathbf{H}' + \hat{\mathbf{V}}_k\right]^{-1} \end{aligned} \tag{10}$$

The matrix inversion lemma in equation (10) reduces the dimensions of the matrices to improve numerical performance. The left-hand term in equation (10) involves inversion of a 4x4 matrix in my simple example, while the right-hand term involves inversion of a 2x2 matrix. When the number of areal controls is larger, say the Census Bureau area for each of 50 counties, and there are 2 remotely sensed strata, say forest and nonforest, then the left-hand term requires inversion of a 100x100 covariance matrix, while the right-hand term inverts a 50x50 matrix. Maybeck (1979, pp. 214-217) uses this lemma, then expands, regroups, and exploits algebraic identities to rewrite equation (9) as:

$$\begin{aligned} \hat{\mathbf{B}} &= \left[\hat{\mathbf{V}}_z - \mathbf{GH}\hat{\mathbf{V}}_z\right] \left[\hat{\mathbf{V}}_z^{-1} \hat{\mathbf{z}} + \mathbf{H}'\hat{\mathbf{V}}_k^{-1} \mathbf{k}\right] \\ &= \hat{\mathbf{z}} + \mathbf{G}\left[\mathbf{k} - \mathbf{H}\hat{\mathbf{z}}\right] \end{aligned} \tag{11}$$

Equation (11) may be expressed in an equivalent “Joseph” form (Maybeck p. 237) as:

$$\hat{\mathbf{B}} = \left[\mathbf{I} - \mathbf{GH}\right] \hat{\mathbf{z}} + \mathbf{Gk} \tag{12}$$

Maybeck (pp. 215-216) derives the covariance matrix for the estimate in equation (11) as:

$$\mathbf{V}_{\hat{\mathbf{B}}} = \hat{\mathbf{V}}_z - \mathbf{GH}\hat{\mathbf{V}}_z \tag{13}$$

An alternative form for equation (13) is:

$$\mathbf{V}_{\hat{\mathbf{B}}} = \left[\mathbf{I} - \mathbf{GH}\right] \mathbf{V}_z \left[\mathbf{I} - \mathbf{GH}\right]' + \mathbf{G}\hat{\mathbf{V}}_k \mathbf{G} \tag{14}$$

While equation (14) is more complex than equation (13), the “Joseph” form of the covariance matrix in equation (14) has numerical advantages (Maybeck 1979, p. 237). However, equation (14) can still have numerical problems with illconditioned covariance matrices. Maybeck (pp. 368-405) develops equivalent estimation equations that improve numerical precision and stability by using square roots of the covariance matrices. Maybeck (1979) recommends diagnostics that test the distribution of multivariate residuals, or “innovations.”

APPLICATION OF GLS TO AREAL CONTROL IN FIA

Let the vector of independent ancillary estimates **k** in equation (5) contain the areal control statistics from the Census Bureau, i.e., **k**=**c** from equation (3). The asso-ciated covariance matrix in equations (10) through (14) equals the zero matrix, i.e., $\mathbf{V}_k=\mathbf{0}$**,** because **c** is a vector of constants. In this case, the usual form of the GLS estimator in equation (8) will not work because it requires the infeasible inverse of a singular covariance matrices. However, the equivalent estimators in equations (11) and (12) are feasible because the weighting matrix **G** in equation (10) simply equals:

$$\begin{aligned}\mathbf{G} &= \hat{\mathbf{V}}_z\mathbf{H}'\left[\mathbf{H}\hat{\mathbf{V}}_z\mathbf{H}' + \mathbf{0}\right]^{-1} \\ &= \hat{\mathbf{V}}_z\mathbf{H}'\left[\mathbf{H}\hat{\mathbf{V}}_z\mathbf{H}'\right]^{-1}\end{aligned} \qquad (15)$$

The matrix inversion lemma in equation (10) assumes that $\mathbf{V}_z$ and $\mathbf{V}_k$ are positive definite covariance matrices (Maybeck 1979, p. 213). This is obviously untrue when $\mathbf{V}_z=\mathbf{0}$ for equation (15), even though equation (15) is feasible. However, this violation of assumptions has never caused any problems for my numerical applications, and an alternative derivation should exist that does not need to assume that $\mathbf{V}_k$ is positive definite. In addition, **y** is a vector of proportions that sum to exactly one, and its covariance matrix $\mathbf{V}_z$ in equation (1) is not positive definite; neither is $\mathbf{HV}_z\mathbf{H'}$ positive definite, which means that its inverse in equation (15) does not exist. This latter problem can be solved by deleting one row in **c** and $\mathbf{HV}_z\mathbf{H'}$, and the corresponding row and column in covariance matrix $\mathbf{HV}_z\mathbf{H'}$, or using the pseudo-inverse.

FIA requires estimators that can be implemented within database structures. This typically requires that statistical estimators be implemented as expansion factors in the database. This criterion cannot be met exactly with the expressions in equations (11) or (12). Equation (11) might be best suited for application in a database. The database retains the expansion factors originally used to estimate vector **z,** which contains the statistics in FIA core tables for area. However, each element of **z**, and each corresponding cell in FIA core tables, must be “adjusted” by the associated element in the vector **G**[**c**-**Hz**], which contains positive and negative elements centered on zero. One way to implement these adjustment terms is to add a “record” to the FIA database that contains the estimated population totals for area in the FIA Inventory Unit (i.e., **z**).

DISCUSSION

I show that there is at least one alternative to stratification for areal control. This alternative might solve problems with insufficient sample sizes in small cross-classified strata, especially as FIA shifts to an annualized system that measures 20 percent or less of its Phase Two field plots each year. In effect, the empirical relationship between remote sensing and field data is established using data across the entire multi-county region to avoid problems with small sample sizes in some strata. This is already captured in equation (1). Then, a linear multivariate estimator in equation (11) or (12) applies areal controls so that final estimates are constrained to agree with “official statistics” (e.g., Bureau of Census).

This same approach can be used to constrain FIA estimates to other “official statistics,” such as the published area for each national forest, ranger district, park, or preserve. If it is not necessary to make FIA estimates agree with cross-classifications of administrative entities (e.g., area of each county on a national forest), then the areal controls can be applied sequentially through the estimators given above.

Because the number of cells in FIA core tables are so numerous, and the GLS estimators require inversion of covariance matrices that correspond to cells in FIA statistical tables, the proposed solution might only work for a few broad indicators (e.g., total forest area), which are typically margins in FIA core tables. The proposed solution might require ad hoc or special methods to implement in FIA information management system.

I plan to test the practical value of these estimators using Monte Carlo simulations.

REFERENCES

Maybeck, P.S. 1979. Stochastic models, estimation, and control. New York: Academic Press. 423 p. Vol. 1.