2 FIM

MIXED ESTIMATION FOR A FOREST SURVEY SAMPLE DESIGN

Francis A. Roesch, USDA Forest Service Southern Research Station
160A Zillicoa Street, P.O. Box 2750, Asheville, NC 28802

Key Words: Forest Sampling, Small-area Estimation.

Abstract: Three methods of estimating the current state of forest attributes over small areas for the USDA Forest Service Southern Research Station's annual forest sampling design are compared. The three methods were (I) simple moving average, (II) single imputation of plot data that had been updated by externally developed models, and (III) local application of a global model that was determined through mixed estimation. In a preliminary analysis of current basal area estimation, the less complex Method III compared favorably with Method II in terms of squared error loss.

Introduction

"...FIA Data are only useful when they are current, consistent and reliable". Rep. Bob Goodlatte, Chairman U.S. House of Representatives Forestry Subcommittee of the Agriculture Committee - The Forestry Source, May 1999

The quotation above reflects the sentiment that had earlier led the USDA Forest Service Southern Research Station (SRS) to initiate an annualized forest inventory sampling design known as SAFIS (Southern Annual Forest Inventory System). SAFIS was introduced in order to improve estimation of both the current resource inventory and changes in the resource. Under the previous periodic inventory system individual states were inventoried over a 1 to 3 year period, about every 10 years. Many factors, including rapid land use change and the intense forest dynamics in the southern United States, contributed to a low amount of confidence in inventory estimates that were more than a few years old. It was decided that an annualized inventory system, in which data is collected statewide every year, would provide more timely and useful estimates.

The sample plots for the SAFIS sample design are located in a systematic triangular grid with five interpenetrating panels. One panel per year is measured for five consecutive years. Every five years the panel measurement sequence reinitiates. If panel 1 was measured in 1998, it will also be measured in 2003, 2008, and so on. Panel 2 would then be measured in 1999, 2004, 2009, etc. The panels are assigned according to the pattern in Figure 1, which results in each element having no immediate neighbors from the same panel.

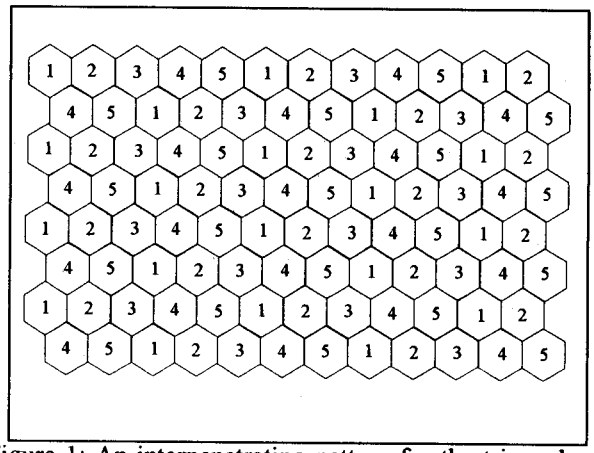

Figure 1: An interpenetrating pattern for the triangular five-panel design. The panels represent each of five consecutive years of measurement.

Analysis

In this paper, I investigate three methods of combining the multi-year data from the SAFIS design to form current estimates for small-areas. I assume that one and only one full series of observations is available. That is, all five panels have been measured once. Current is defined as the time of measurement of panel 5. The estimation system for a production inventory should be judged by how well it balances statistical efficiency with simplicity and the ability to be implemented in an unsupervised program.

Estimation Methods

Method I pools the latest 5 panels measured, rendering it equivalent to a five-year moving average, when applied yearly. For a single instance, this is similar to the method currently used by FIA for periodic inventories in states that required more than one year to inventory. The five-year moving average has been declared the default estimator by FIA (Roesch and Reams, 1999). It is obtained by assuming that there is no time trend at the observed scale, since it will perform poorly in the presence of monotonic trend. A practical advantage to using the moving average approach, initially, is that the currently existing software will be applicable.

Because the time duration of measuring all five panels is somewhat longer than the duration of 1 to 3 years it took for the periodic inventories, equal

Reprinted from the 1999 Proceedings of the
Section on Statistics and the Environment of the American Statistical Association

weighting of panels may have the tendency to mask the trends that the SAFIS design was intended to evaluate. Therefore, rather than assuming that there is no time trend, we might favor methods which recognize and efficiently utilize the time-series nature of the five-panel sample. Methods II and III do this in different ways. Method II uses externally developed equations to project the plot basal areas measured in years 1 through 4 to year 5. Method III attempts to model the time trend of plots within a panel series. A mixed estimator proposed by Van Deusen (in review) is used that can incorporate increasing levels of constraints on the derivatives of the time trend, allowing one to model various levels of complexity in the time trend. The mixed estimator literally mixes two models, the first describes the relationship of observations within each panel (or time period) and the second describes constraints on the time trend. The mixed estimation approach is expected to be both powerful and practical for most variables of interest to FIA, and performed well in the case study described below.

Often forestland is subdivided into mutually exclusive condition classes, which are observed on the sample plots. A condition class is at least an acre in size and is classified by land use, forest type, stand origin, stand size, stand density, and ownership class (Anonymous, 1998). The usual focus is on estimation of the current per acre value (V) of an attribute for a particular condition class k. Let:

$p_{i(j)}$ = plot i within county j ($i = 1,\ldots,n_j$),

o_j = county j ($j = 1,\ldots,J$),

t_t = time t ($t = 1,\ldots,5$),

c_k = condition class k ($k = 1,\ldots,K$),

$X_{i(j)tk}$ = the per acre value observed at $p_{i(j)}$, o_j, and t_t, for c_k,

$A_{i(j)tk}$ = the area in acres sampled in c_k at $p_{i(j)}$, o_j, and t_t

$$C_{i(j)tk} = \begin{cases} 1 & \text{if } c_k \text{ occurs at } p_{i(j)},\ o_j, \text{ and } t_t, \\ 0 & \text{Otherwise} \end{cases}$$

A_P = Plot area

In the case where there is no time trend present, the overall mean for the five panel series would provide the best estimator of a per acre value (V) for condition class k:

$$V_k = \sum_{t=T-4}^{T} \sum_{j=1}^{J} \frac{1}{A_{jtk}} \sum_{i=1}^{n_j} \frac{A_{i(j)tk}}{A_P} X_{i(j)tk} \qquad (1)$$

where:

A_{jtk} = Sum of the plot areas sampled in condition class k, in county j at time t.

Equation (1) pools estimates from latest 5 panels measured and, when applied yearly is equivalent to a five-year moving average. The value over all condition classes is simply estimated by:

$$V_{MA} = \left[\left(\sum_{k=1}^{n_k} V_k A_{S_k} \right) \Big/ \left(\sum_{k=1}^{n_k} A_{S_k} \right) \right] \qquad (2)$$

where:

$$A_{S_k} = \sum_{t=1}^{T} \sum_{j=1}^{J} A_{jtk}$$

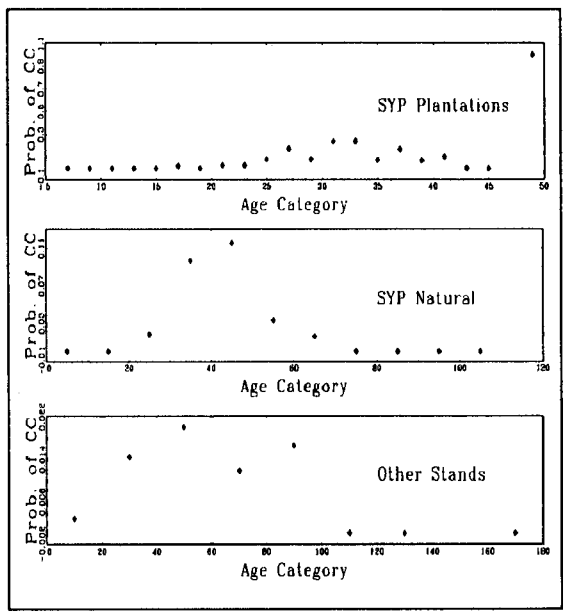

Figure 2: Probability of clear felling by broad forest type and age category.

Method II, favored heavily in industrial forest inventories, utilizes a compilation of externally developed growth and mortality models to project the basal areas of the plots measured at times 1 through 4 to time 5. Proprietary software which implements results of growth and yield studies conducted at Virginia Polytechnic Institute & State University and the University of Georgia was provided by Champion International Corporation for the purposes of this study. This leaves either the prediction or observation of harvest in order to update the basal area estimates. Two cases were investigated. In case 1 it is assumed that harvests were not observed, requiring that both full and partial harvests be modeled (For example the probability of clear felling by broad forest type and age category given in Figure 2 could be randomly applied to each

plot). In case 2, it is assumed that clear-cut harvests are known, leaving only partial harvests to be modeled.

Method III involves two variations of an application of mixed estimation to the SAFIS design discussed in Van Deusen (in review). A brief explanation of Van Deusen's method appears below. Each variation applies global (survey unit) results of the mixed estimation methodology below the survey unit level, under the assumption that the sample will often be too small for a direct application of mixed estimation below the survey unit level. In both variations, mixed estimation was used at the survey unit level to choose from the three simple models discussed by Van Deusen (in review), a flat line, a straight line and a quadratic, and to find the maximum likelihood estimate of the weighting parameter p. In variation 1, the model and the appropriate level of p were then fit at lower levels (i.e. county and county group). In variation 2 the model was fit at the survey unit level to predict an overall $\hat{\beta}$ (a 5x1 vector described below). This leads directly to a simple updating vector $\mathbf{U}$ found by multiplying the inverse of each element of $\hat{\beta}$ by the fifth element of $\hat{\beta}$. Then:

$$V_{MX2} = \left((\mathbf{1}'\mathbf{A}_T)^{-1} ((\mathrm{DIAGRV}(\mathbf{A}_T))\mathbf{V}_T)' \mathbf{U} \right)$$

where:

$\mathbf{A}_T$ = a Tx1 vector of total area sampled at each time,

$\mathbf{V}_T$ = a Tx1 vector of basal area estimates for each time,

$\mathbf{1}$ = a Tx1 vector of ones, and

$\mathrm{DIAGRV}(\mathbf{A}_T)$ is a function that places a Tx1 vector $\mathbf{A}$ into the diagonal of a TxT matrix of zeroes.

Van Deusen's mixed estimator:

Van Deusen (in review, referencing Theil, 1971) proposed the following mixed estimator. First a simple model for the sample data at time $t=1,\ldots,T$ is used:

$$\bar{y}_t = \beta_t + \bar{e}_t \qquad (3)$$

where β_t is an unknown coefficient, $\bar{e}_t$ is an error term with a mean of 0 and a variance of σ_t^2/n_t. We would estimate the error term by the usual sample estimator. Collect the $\bar{y}_t$'s into the vector $\overline{\mathbf{Y}} = [\bar{y}_1,\ldots,\bar{y}_T]'$, and the error terms into the vectors $\mathbf{e} = [\bar{e}_1,\ldots,\bar{e}_T]'$. The matrix representation of equation (3) is then:

$$\overline{\mathbf{Y}} = \beta + \mathbf{e}. \qquad (4)$$

where β is a Tx1 vector of β_t's. Represent the covariance matrix of $\overline{\mathbf{Y}}$ with Σ. Constraints on the time progression of β_t are accounted for in a second model:

$$\mathbf{R}\beta = \mathbf{v} \qquad (5)$$

where $\mathbf{R}$ is an appropriately sized matrix of constraints and $\mathbf{v}$ is an error vector of zero mean and $p\Omega$ variance. Combine equations (4) and (5) into:

$$\begin{bmatrix} \overline{\mathbf{Y}} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{R} \end{bmatrix} \beta + \begin{bmatrix} \mathbf{e} \\ \mathbf{v} \end{bmatrix}$$

The mixed estimator is:

$$\hat{\beta} = \left(\sum^{-1} + \frac{1}{p}\mathbf{R}'\Omega^{-1}\mathbf{R} \right)^{-1} \left(\sum^{-1} \overline{\mathbf{Y}} \right)$$

and the covariance matrix:

$$V(\hat{\beta}) = \left(\sum^{-1} + \frac{1}{p}\mathbf{R}'\Omega^{-1}\mathbf{R} \right)^{-1}$$

Methods

The data from FIA's survey unit 1 in Georgia were used to simulate a data set that might have been obtained had the SAFIS design been in place. The survey unit consists of 35 counties, which were grouped into five contiguous 7-county groups for part of this study. The measurements from 1989 and 1996 were used to simulate the data set that might have been obtained if the SAFIS design had been initiated in 1994 and all measurements through 1998 had been completed. Individual tree basal area projection equations, mortality and harvest probabilities and proportions were established from horizontal point sample clusters measured in 1989 and 1996. These functions were then applied to the 1996 tree level data from fixed-area plot clusters that were co-located with the point sample clusters to project them backward 1 and 2 years and forward 1 and 2 years. This resulted in simulated tree data for five consecutive years on 2,342 fixed-area plots. This data set was considered the "truth" for each of the years 1994 through 1998. The "true" mean basal areas per acre, by year, for the survey unit are graphed in Figure 3. The "current truth" was defined as the values of this data set for 1998.

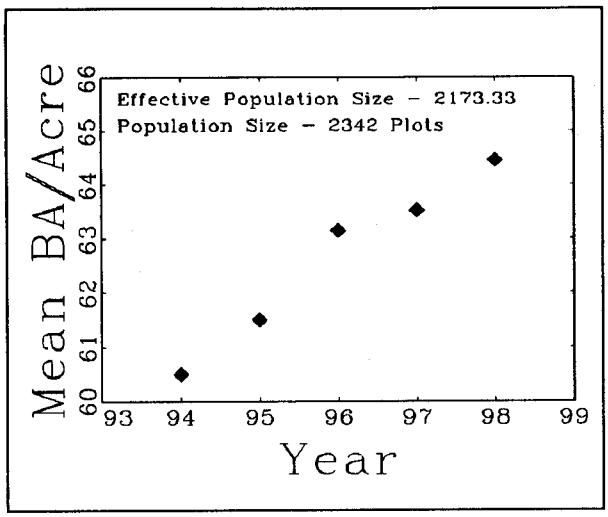

Figure 3: Survey Unit "True" mean Basal Area Per Acre by year.

To effect the systematic SAFIS sample design, spatial coordinates of the plots were used to assign plots to panels, a panel being a single year's measurement. Therefore the simulated SAFIS sample consisted of approximately 1/5 of the plots for each year. Figure 4 depicts the sample mean basal area per acre by year for the survey unit.

The estimators described above were then evaluated for how well they estimated the "true" county level and county group level basal areas for 1998 from the 1994 to 1998 sample, under a squared error loss function, in a preliminary case study. Because this was a case study there was a unique solution for the moving average estimator (Method I) and each variation of Method III. The squared error calculated for these methods is simply the mean of the squared difference of each estimate by county and county group from the truth for that county or county group. Since the two cases of Method II applied random harvests (Case 1 to all harvests and Case 2 to partial cuts), these were simulated 1,000 times and the squared error calculated is actually a mean of the squared error over the 1,000 simulations.

Results

Figure 5 graphs each basal area per acre estimate relative to the "truth" by county (upper graph) and county group (lower graph). It is noteworthy that in both graphs all of the values for the moving average are below the line as a result of the increasing trend in the variable of interest. In addition, the most widely varying estimator at the county level is variation 1 of the mixed estimation approach. This appears to be due to the fact that the sample sizes are too small at the county level to fit the model. This statement is supported by two observations: (1) variation 1 is better behaved at the county group level, and (2) variation 2, in which the model was fit at the survey unit level and then applied at the lower levels, works well even at the county level.

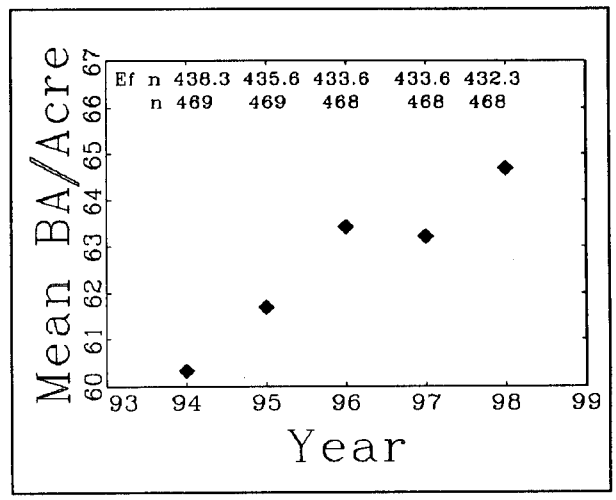

Figure 4: Survey Unit sample mean Basal Area Per Acre by year.

The upper graph in figure 6 shows the mean difference from the truth over all counties and county groups for all three Methods. The lower graph shows the corresponding mean squared differences. The mean squared differences for the panel 5 mean and mixed estimation variation 1 are shown in tabular form to enhance clarity for the other estimators. The panel 5 mean is included this estimator utilizes only the sample data observed on the population of interest (that is basal areas for 1998). Because this is a case study, the mean difference graph is not a reliable measure of model bias but it might give us an indication of model bias. Note that four of the estimators have roughly the same mean difference at both the county and county group levels, leading one to suspect that the respective levels may be reflective of the true level of bias in the model. Of these four, the moving average shows the largest absolute difference, while Method II when clear felled harvest areas are not known comes in second at just over half of the magnitude. When clear felled areas are known, the magnitude of the absolute mean difference is very close to zero for Method II, as it is for the second variation of Method III. The large reduction in magnitude of absolute mean difference for the other two estimators when going from the county to the county group level, suggests that the large absolute mean difference may be more a result of the large variance at that level than of

bias. Of course, we know for one of these estimators, the panel 5 mean, that this is the case, because the panel 5 mean is design unbiased and does not rely on a temporal model. The two estimators which show the lowest mean squared differences overall are the basal area projection model when clear felled harvest areas are known, and variation 2 of the mixed estimation method.

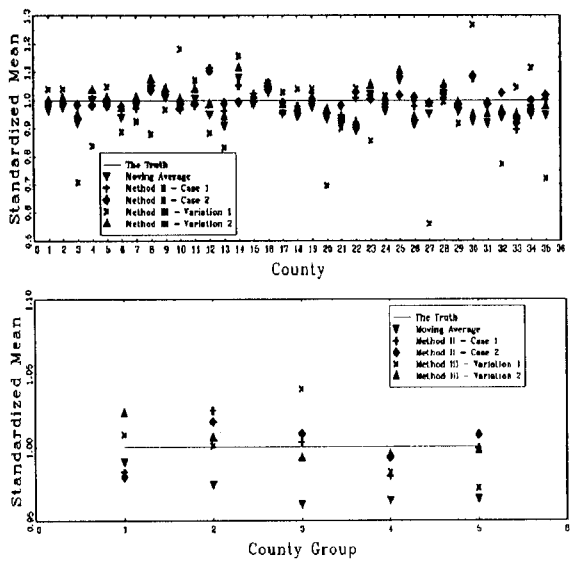

Figure 5: County (upper) and County Group (lower) means relative to the "true" population mean for (1) Simple moving average, (2) Single imputation of projections with modeled clear felled areas, (3) Single imputation of projections with known clear felled areas, (4) Mixed Estimator Variation 1 – Model and P selected at survey unit level and fit at lower levels, (5) Mixed Estimator Variation 2 – Model and P selected and fit at survey unit level and applied at lower levels.

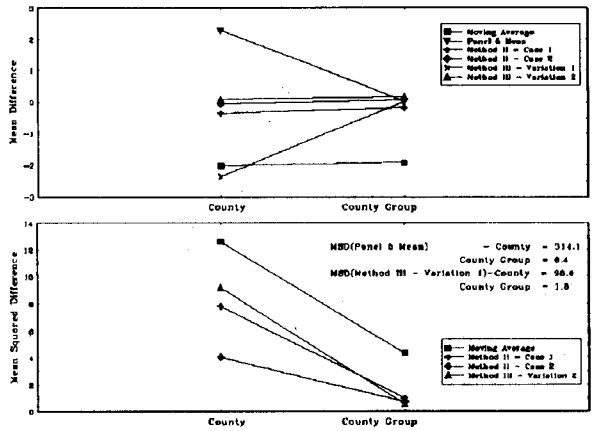

Figure 6: Mean sum of differences (upper) and mean sum of squared differences (lower) between each estimator and the "truth" over all counties (left) and county groups (right) for (1) Simple moving average, (2) Single imputation of projections with modeled clear felled areas, (3) Single imputation of projections with known clear felled areas, (4) Mixed Estimator Variation 1 – Model and P selected at survey unit level and fit at lower levels, (5) Mixed Estimator Variation 2 – Model and P selected and fit at survey unit level and applied at lower levels.

Conclusions

Methods I, II, and III all use outside information in some sense. All three methods resulted in a substantial improvement in terms of squared error loss over the single panel mean. None of these alternative estimators, as applied to the small-areas, however, can be shown to be design unbiased. All of the alternative estimators, except for the simple moving average in the presence of monotonic trend, have the potential of being model unbiased. For basal area (and presumably all variables that are likely to exhibit trend over the 5-year measurement period) even simplistic approaches to modeling the trend can result in significant reductions in squared error over the simple moving average.

In comparing Methods II and III, I note that case 1 of Method II is the proper case to compare to Method III since case 2 of Method II utilized information not provided to Method III. That information was knowledge of the cleared-felled areas. It's quite possible that Method III, variation 2 would have done just as well at the county level if it had been applied to cleared and non-cleared areas separately. Data on cleared and non-cleared areas are not currently collected annually, and I merely intended to demonstrate the statistical advantage of doing so, since clear-felled areas have a distinctly different basal area than non clear-felled areas. The former having basal areas of zero or close to zero.

Method III, in general, represents a much lower investment in human resources both initially and in the long term than Method II. Although Method II appeared to work well in this case study, it is true that appropriate growth models do not exist for many condition classes of interest, and those that do exist would have to undergo thorough testing in this context. In addition, the growth model predictions would have to be constantly monitored to ensure that the forest populations are not moving away from those upon which the models were built, thereby reducing the reliability of the models.

The success of Method II depends on how well basal area development in the target population of small areas matches basal area development in previously measured populations. The success of Method III depends on how well basal area development in the individual small areas matches basal area development in the large area (survey unit). Method III could have been expected to perform well for basal area and other variables of relatively low variance and high frequency of observation. The size of the small area would have to be increased for attributes that are more rarely observed or more highly variable. Method II, on the other hand, which inherently incorporates a broader information base, might be expected to perform well for these more difficult to estimate attributes.

A problem with case studies is that the population is fixed and whatever methods were used to construct the population predetermine the results of estimator comparisons. Therefore a method might be inadvertently favored in a case study that would not fare as well in a more realistically robust set of populations. Ergo, an obvious future direction for this work would be to simulate a set of populations that can be assumed to be realistically robust to further test the competing estimation systems.

Literature Cited

Anonymous. 1998. *Field Instructions for Southern Forest Inventory, version #2.* Southern Research Station. USDA Forest Service, Asheville, NC.

Roesch, F. A. and G. A. Reams. 1999. Implementation of an Annual Inventory Design. *Jour. For.* 97(12):xxx-xxx.

Theil, H. 1971. *Principles of Econometrics.* John Wiley & Sons. New York.

Van Deusen, P. C. In review. Modeling trends with annual survey data.