Alternative Sampling Designs and Estimators for Annual Surveys

Paul C. Van Deusen

Abstract.—Annual forest inventory systems in the United States have generally converged on sampling designs that: (1) measure equal proportions of the total number of plots each year, and (2) call for the plots to be systematically dispersed. However, there will inevitably be a need to deviate from the basic design to respond to special requests, natural disasters, and budgetary fluctuations. Some simple methods for maintaining a flexible panel design and implementing a moving average estimator are discussed.

There are several practical and theoretical reasons why annual systems have recently gained support. Equal annual budgets and work loads are easier to support politically than the fluctuations that result from the traditional periodic approach. Field crew retention improves when there is assurance of regular work within a compact region. Cooperation between the appropriate federal and state agencies can be regularly fostered without the multiyear hiatus that occurred under the old system. Statistically, there is an opportunity to realize gains from incorporating models and time series analysis that didn't exist under the old system. Something analogous to double sampling (Fairweather and Turner 1983, Hansen 1990) is very natural with an annual system. It's natural to take the expensive measurements on some of the plots each year while simulating measurements on the remaining plots. If the simulated measurements are "good," then overall precision is improved or fewer measurements need to be taken. Trends can be detected with only a 1-year lag under an annual system, as opposed to a multiyear lag under the traditional system. Flexibility is generally greater under an annual system. When budgets decrease, the proportion of plots can be reduced. Likewise, increasing budgets can be accommodated by measuring a larger proportion of plots.

Two well-established annual forest inventory systems are being tested and implemented at this time by the USDA Forest Service. The original system, AFIS, was developed in Minnesota, and its younger cousin (SAFIS) is being implemented throughout the South. SAFIS was conceived as a simplification of AFIS based on systematically measuring an equal proportion of plots each year. The AFIS designers originally focused on keeping costs down and felt that disturbance-based plot sampling would yield efficiencies that would justify its theoretical complexity. As public support has grown for AFIS and SAFIS, the need for cost savings has diminished relative to the need for producing publicly accessible methods and data. However, circumstances will inevitably occur where the basic SAFIS design will need to be complicated, at least temporarily.

Natural disasters such as hurricanes or ice storms that affect large forested areas are usually followed by a demand for a special inventory to assess damage. Budgetary decreases may make it impossible to complete field work as planned, while increases may require doing more than originally planned. There is also a need for SAFIS and AFIS to converge on similar sampling designs and analysis procedures as annual systems begin to be implemented nationally. The Forest Service receives regular feedback from the public on the need for compatible inventory procedures across regions and plans to accommodate these requests. Therefore, methods should be developed for enhancing the original SAFIS design or, depending on your point of view, simplifying the AFIS design.

DESIGN MODIFICATIONS

Only some very simple variations on the basic sample design will be discussed. A rigid SAFIS design would measure an equal proportion of plots, say 20 percent, each year. One way to do this would be to allocate each plot to one of five panels, and then measure one panel per year. Therefore, each panel would be regularly measured at 5-year intervals. The problem with this approach is that a pure non-overlapping panel design will only have data from which to estimate 5-year growth intervals. Therefore, other estimates will depend on the models, since plot updates will be over intervals that are less than 5 years. Such estimates are not design-unbiased. It's worth noting that a pure moving average estimator would be practically design unbiased even with rigid panels. The good news is that this may not be a problem for at least 5 years, because plots will have different measurement histories when they first enter the annual system. For example, the State of Georgia had plots that had been measured from 1 to 4 years earlier when they began the annual system in 1998.

Principal Research Scientist, NCASI, Anderson Hall, Tufts University, Medford, MA, USA, Pvandeus@tufts.edu

A possible solution is to follow the basic SAFIS plan for 5 years. After that, simply reshuffle the panel assignments in such a way that no plot is remeasured more frequently than 3 years or less frequently than 10 years. This would still result in five panels and would be transparent to the users. It allows use of a moving average estimator and resolves the concern about having only 5-year growth intervals. It would make model calibration easier as well. To implement the reshuffling, one might consider creating clusters of five neighboring plots each; then do the reshuffling within each cluster to maintain systematic coverage in each panel.

The rigid SAFIS design can't deal with budgetary fluctuations with a five-panel system. This issue could be handled by creating "extra" panels. Suppose you create 10 panels instead of 5. As long as there are enough resources to measure 20 percent each year, measure 2 of the 10 panels per year, i.e., the extra panel labels don't serve any purpose. However, if the budget starts to fluctuate, consider only measuring one panel or increase to measuring three panels. This maintains a panel design that systematically covers the region. The "extra" panel concept is a device to allow more flexibility in the future or for states that don't need 20 percent coverage each year. One could implement and maintain a 10-panel system by creating clusters of 10 plots each. The panel assignments within the clusters could be reshuffled periodically to maintain a mix of measurement intervals as well.

ANALYSIS OPTIONS

Much has been written elsewhere about analysis options (Hansen 1990, Reams and Van Deusen 1998, Van Deusen 1996a). These options can be grossly categorized into two groups, those that require imputed (modeled) values and those that don't. Imputation methods include single imputation and multiple imputation (Rubin 1987, Van Deusen 1996a). Multiple imputation has the advantage of allowing for simple estimates of means and variances, while single imputation leads to complicated variance estimates and the temptation to use ad hoc approaches. Regardless, multiple imputation is no panacea, because the bias and variance characteristics of these estimates will ultimately depend on the quality of the imputed values.

Two imputation-free methods can be applied to annual systems. The first is based on the fact that each panel provides an independent random sample from the entire population. Therefore, each panel can produce estimates in the year it was measured independent of the other panels. The second method involves combining panels measured in different years using a moving average or time series methodology. Therefore, the second method might be called the moving average approach, but it actually encompasses a number of procedures. For example, one might consider variations on the Kalman filter (Kalman 1963), mixed estimators (Theil 1971), exponentially weighted moving average (EWMA) models, etc. Since this is likely to be an important analysis method for annual systems, I will present the details for a mixed estimation approach.

Mixed Estimator

The mixed estimator (Theil 1971) has certain advantages over the Kalman filter or ad hoc moving average approaches in the context of annual surveys. The Kalman filter tends to have a start-up period of several years where the influence of initial starting values may dominate the data. Likewise, moving average methods often call for pre-specified weights to be applied to each of the input data values. The mixed estimator has neither of these deficiencies, although it is closely related to the Kalman filter.

The proposed mixed estimator is based on an observation equation that relates the observed panel means to the unknown population mean,

$$y_t = \beta_t + e_t \qquad t=1,\ldots,T \qquad (1)$$

where y_t is the mean from the panel measured in year t and e_t is an error term with mean 0 and variance σ^2_t/n_t. Model (1) by itself leads to estimating the population mean from the sample mean at time t. The population variance, σ^2_t, can be estimated in the usual way from the sample data. A second model is required to describe the transition of the mean over time. The transition equation could also be used to incorporate imputed values (Van Deusen 1996b). The simplest transition equation results from the assumption that the mean follows a random walk over time,

$$\beta_t = \beta_{t-1} + v_t \qquad t=2,\ldots,T \qquad (2)$$

where v_t is an error term with mean 0, variance $p\sigma^2_t/n_t$ and T corresponds to the cycle length.

The mixed estimator (Theil 1971) provides a method for combining equations (1) and (2) to produce estimates of the population mean for times t=1,...,T along with a covariance matrix. More elaborate notation is required to allow for plots being measured for the first time or remeasured at any sample occasion. Sampling with partial replacement methods (Van Deusen 1989) can serve as a template for adding any sequence of remeasurements.

Equation (2) allows the true mean to vary quite freely over time. Therefore, a more restrictive transition equation might be in order such as,

$$\beta_t = 2\beta_{t-1} - \beta_{t-2} + v_t \qquad t=3,\ldots,T \quad (3)$$

Equation (2) is controlling the slope of the mean transition trend, whereas equation (3) controls the second derivative of the mean trend, which will lead to smoother transitions.

Matrix notation facilitates derivation of estimators. Let $Y=[y_1,\ldots,y_T]'$, $\mathbf{B}=[\beta_1,\ldots,\beta_T]'$ and $\mathbf{e}=[e_1,\ldots,e_T]'$. Then equation (1) can be written in matrix form as $\mathbf{Y}=\mathbf{B}+\mathbf{e}$. To do the same for equation (2), let $\mathbf{R}$ be a T-1 by T matrix of the form

$$R = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ & & & & \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix} \quad (4)$$

and let $\mathbf{v}=[v_2,\ldots,v_T]'$. Now equation (2) can be simply written in matrix form as $\mathbf{0}=\mathbf{RB}+\mathbf{v}$, where $\mathbf{R}$ performs the task of first-differencing the parameters in $\mathbf{B}$. Equation (3) can be accommodated by defining $\mathbf{R}$ as T-2 by T with each row containing the sequence (1 -2 1) and eliminating v_2 from $\mathbf{v}$. Now the observation and transition equations can be written in matrix form as

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} D \\ R \end{bmatrix} B + \begin{bmatrix} e \\ v \end{bmatrix} \quad (5)$$

where $\mathbf{D}$ is a design matrix of 0's and 1's that serves to match up the elements in $\mathbf{Y}$ with the appropriate parameter in $\mathbf{B}$. For the example being discussed, D is an identity matrix.

Now define $\Sigma=\mathrm{diag}(\sigma^2_1/n_1,\ldots,\sigma^2_T/n_T)$, the covariance matrix for the error vector $\mathbf{e}$. More generally, Σ would contain off-diagonal terms to account for covariances between remeasured plots. Likewise, the covariance matrix for $\mathbf{v}$ is $p\Omega$, where Ω is a diagonal submatrix of Σ. Using this notation, the mixed estimator for $\mathbf{B}$ (Theil 1971) is

$$\hat{B} = \left(D'\Sigma^{-1}D + \frac{R'\Omega^{-1}R}{p} \right)^{-1} D'\Sigma^{-1}Y \quad (6)$$

with covariance matrix

$$V(\hat{B}) = \left(D'\Sigma^{-1}D + \frac{R'\Omega^{-1}R}{p} \right)^{-1} \quad (7)$$

The diagonal elements of $V(\hat{\mathbf{B}})$ provide the estimated variances for the population mean estimates for times t=1,...,T. As p increases, the influence of the transition equation diminishes. In fact, as $p \to \infty$ the mixed estimator is the same as obtaining independent estimates from each panel when Σ is diagonal. Confidence intervals around change estimates are also available since $V(\hat{\mathbf{B}})$ contains the necessary covariance terms. The assumption here is that finite population correction factors (FPCs) can be ignored, because forest sampling usually involves large total area relative to the sampled area.

Maximum Likelihood Estimation of Unknowns

The p-parameter can be estimated along with $\mathbf{B}$ by maximizing the appropriate joint distribution. Assuming that the error vectors, $\mathbf{e}$ and $\mathbf{v}$, from the observation and transition equations follow independent multivariate normal distributions, their joint likelihood can be written as

$$L \propto \frac{1}{2}[\log(|\Sigma|) + (Y-B)'\Sigma^{-1}(Y-B) + \frac{1}{p}B'R'\Omega^{-1}RB + \log(|p\Omega|)] \quad (8)$$

The maximum likelihood estimates are those that minimize L. Given our assumptions, everything is known in equation (8) except p. For the example application, no effort is made to find the maximum likelihood estimate for p.

EXAMPLE APPLICATION

The purpose of the example is to demonstrate the effect of the transition equations and the p-parameter with all else held constant. A data set was simulated for 5 years with annual means of 10, 30, 15, 18, and 22. The diagonal terms in Σ and Ω were all set to 1. Equation (6) was used to fit the mixed estimator to these data with the variance parameter set to p=1 under transition equation (2) (fig. 1a) and for transition equation (3) (fig. 1b). The analysis is repeated (fig. 2) with p=5 to demonstrate how larger p-values increase the flexibility of the model. Comparison of results from the two transition equations with common p shows that equation (3) leads to smoother trends. In a real application, the components of Σ would be estimated from the data and p should be the value that maximizes likelihood function (8).

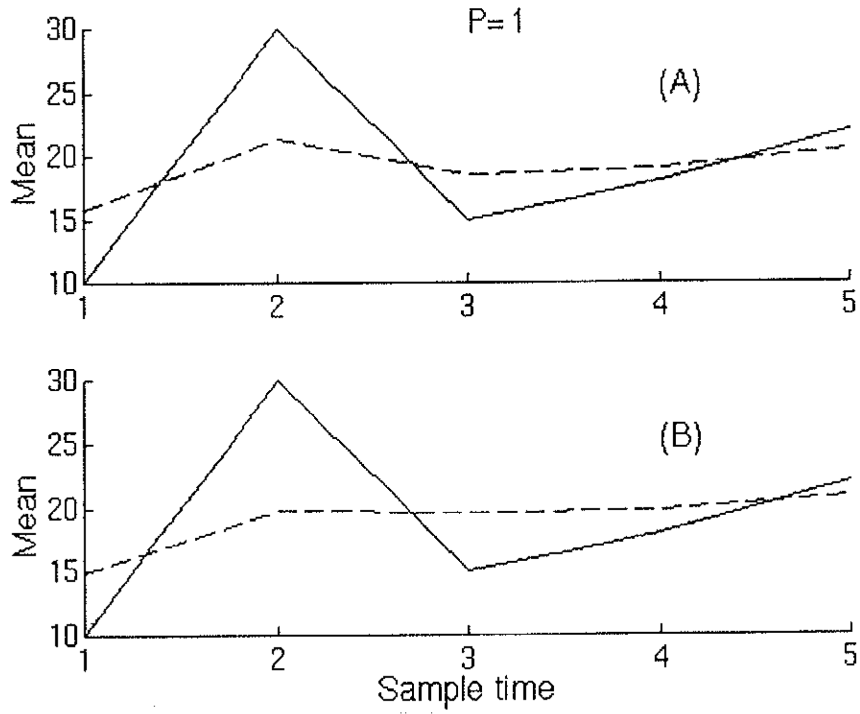

Figure 1.—*The mixed estimator (dash line) using (A) transition equation (2) and (B) transition equation (3) with p=1.*

CONCLUSIONS

Annual inventory systems based on systematic annual samples can provide timely and theoretically valid sample estimates. Minor variations from a rigid panel design have been discussed that can ensure that resulting estimates are practically design-unbiased. Estimation procedures based on imputation will likely come into use

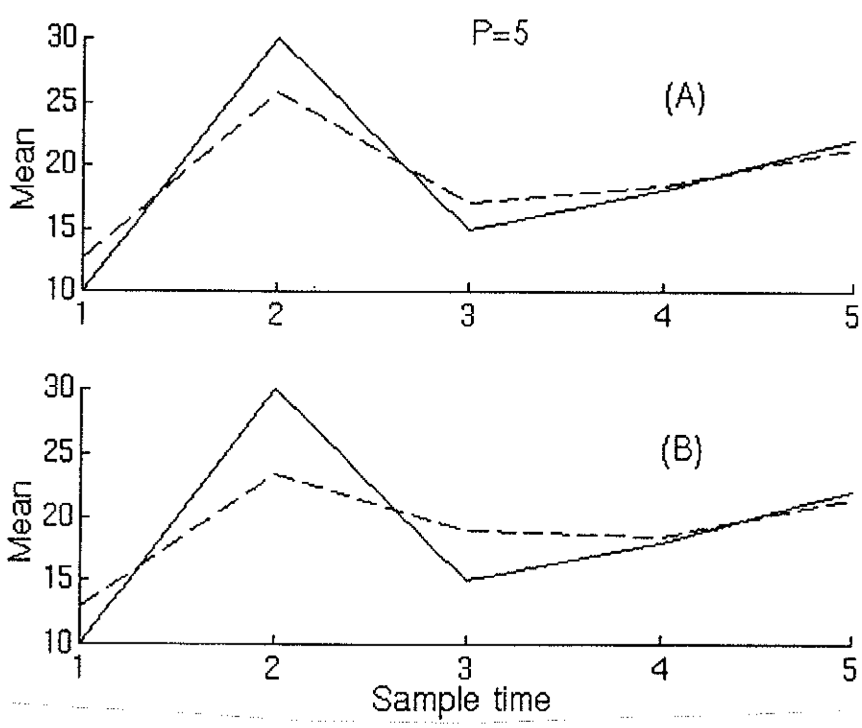

Figure 2.—*The mixed estimator (dash line) using (A) transition equation (2), and (B) transition equation (3) with p=5.*

after sufficient research is done to validate resulting estimates. As an alternative, variations of moving average estimators can provide practically design-unbiased estimates with well-defined variance estimators. In particular, a mixed estimator was developed here that is easy to apply.

ACKNOWLEDGMENTS

The following people reviewed this paper: Dr. Gregory A. Reams, USDA Forest Service, Asheville, NC, USA, and Dr. Francis A. Roesch, USDA Forest Service, Asheville, NC, USA.

LITERATURE CITED

Fairweather, S.E.; Turner, B.J. 1983. The use of simulated remeasurements in double sampling for successive forest inventory. In: Bell, J.F.; Atterbury, T., eds. Proceedings, Renewable resource inventories for monitoring changes and trends; 1983 August 15-19; Corvallis, OR: 609-612.

Hansen, M.H. 1990. A comprehensive sampling system for forest inventory based on an individual tree growth model. St. Paul, MN: University of Minnesota. Ph.D. dissertation.

Kalman, R.E. (1963). New methods in Wiener filtering theory. In: Bogdanoff, J.L.; Kozin, F., eds. Proceedings of the First symposium on engineering applications of random function theory and probability. Wiley: 270-388.

Reams, G.A.; Van Deusen, P.C. 1999. The southern annual forest inventory system.Journal of biology, Agriculture, and Environment Statistics.

Rubin, D.B. 1987. Multiple imputation for nonresponse in surveys. New York, NY: Wiley.

Theil, H. 1971. Principles of econometrics. New York, NY: Wiley.

Van Deusen, P.C. 1989. Multiple-occasion partial replacement sampling for growth components. Forest Science. 35(2): 388-400.

Van Deusen. P.C. 1996a. Annual forest inventory statistical concepts with emphasis on multiple imputation. Canadian Journal of Forest Research. 27: 379-384.

Van Deusen. P.C. 1996b. Incorporating predictions into an annual forest inventory. Canadian Journal of Forest Research. 26: 1709-1713.