Pseudo-CFI for Industrial Forest Inventories

Francis A. Roesch

Abstract.-Corporate inventory systems have historically had a greater spatial and temporal intensity than is common in the public sector. For many corporations, these inventory systems might be described as dynamic in that current estimates rely on a small amount of recent data and a large amount of information resulting from the imputation of older data that have been subjected to various growth and mortality models. Usually the "best available" models are used for this purpose, with little attention paid to any population dynamics that may have occurred since development of the models. This paper gives the theory and an example application of a family of sample designs that possess continuous forest inventory (CFI) attributes. This family of Pseudo-CFI sample designs was devised to facilitate the incorporation of a continuous monitoring and calibration mechanism for the imputed data.

It is important to industrial forest enterprises to know what is and what will be on the ground (by a wide array of measures) at any given point in time. For this reason, corporate inventory systems have historically had a greater spatial and temporal intensity than is common in the public sector. For many corporations, these systems might be described as dynamic in that current estimates rely on a small amount of recent data and a large amount of imputed data values based on older data that have been subjected to a growth model. Usually the "best available" growth and yield models are used for this purpose, with little attention paid to the appropriateness of the models for specific applications. Unfortunately, industrial forest populations are themselves quite dynamic and ever more frequently prove to be quite different from any of those upon which existing models were built. Furthermore, we can expect this trend to continue indefinitely. Therefore, it would be prudent to incorporate a continuous monitoring and calibration mechanism into the inventory system in order to provide the ability to adapt to changing conditions and populations. This paper presents a family of sample designs, which lie between periodic inventories and continuous forest inventories with respect to the variance and cost scales. An example application is also included. Because the designs are obtained by relaxing the requirements of continuous forest inventory, these designs might be dubbed Pseudo-CFI.

Much work has been devoted to studying the continuum along which different methods of scientific inquiry could be positioned with respect to the level of control over responses (Basu 1980; Cochran 1965; Cochran and Rubin 1974; Rubin 1973a, 1973b, 1974, 1976a, 1976b, 1978, 1979, 1980, 1986; Wickramarante and Holford 1987; Wold 1956). The subject was briefly introduced into the forestry literature in a discussion of forest response studies by Green and Roesch (1993). A major premise of this paper is that there is a point at which observational forest inventory data can begin to overrule the results of controlled growth and yield experiments. Here, we show how one might exploit that premise in the design of a dynamic inventory system.

Arguments for strict experimental control in growth and yield estimation and for the use of the less controlled continuous forest inventory (CFI) approach both rely on a very basic truth about random variables. That truth is that the variance of the difference between two random variables (A and B) is:

$$Var(A-B)=Var(A)+Var(B)-2\,Cov(A,B)$$

If we let A equal time 2 volume and B equal time 1 volume, we see that A-B equals volume growth. Both controlled experiments and CFI plots attempt to maximize the third term on the right-hand side (rhs) of the equation, $2Cov(A,B)$, by measuring time 1 and time 2 volumes on the same population elements. This results in a decrease, of course, on the left-hand side (lhs) of the equation. Controlled experiments go one step further by also reducing the first two terms on the rhs. This is accomplished by controlling the population elements entering the experiment to a particular subset of the population of interest. That is, the point of an experiment is to deduce the effect of an action, and in order to do that one tries to eliminate potential noise factors.

A potential weakness of the experimental design approach is the assumption that an effect measured on a very controlled subset of the population is going to be the same as the effect on the entire population. Unfortunately,

Mathematical Statistician, USDA Forest Service, Forest Inventory and Analysis, Southern Research Station, PO Box 2680, Asheville, NC 28802, USA.

noise is reality, and when one eliminates noise, one may also be eliminating major segments of the population of interest and their corresponding differences in response.

This realization in other industries has given rise to such areas of study as process control, updating, and calibration techniques. In forestry, the term "localizing" is often used in reference to adapting volume and/or growth and yield functions to local data. Variations of this approach have been labeled "adaptive growth modeling" by some.

PSEUDO-CFI

Let's consider a continuum of control and information with respect to growth and yield studies. We would observe the following relationships among conditional covariances between time 1 (*T1*) and time 2 (*T2*) data,

$$Cov\left(T_2, T_1 \middle| PI\right) Cov\left(T_2, T_1 \middle| CFI\right) Cov\left(T_2, T_1 \middle| EGP\right)$$

where:

PI = periodic inventories
CFI = continuous forest inventories
EGP = experimental growth plots

Similarly, the conditional variances would be ordered,

$$Var\left(T_2 \middle| T_1, PI\right) Var\left(T_2 \middle| T_1, CFI\right) Var\left(T_2 \middle| T_1, EGP\right).$$

Naturally, the advantage of the higher variance methods is that the costs are ordered:

$$Cost(PI)\ Cost(CFI)\ Cost(EGP)$$

Realistically, the major driver of the level of effort in industrial forest inventories is cost. What are needed are lowest variance estimates obtainable for a given budget. Because some efforts to increase the covariance between time 1 and time 2 attributes will be inexpensive relative to the amount of increase, while others will not, it is worthwhile to seek an optimal point along the covariance/cost function. Note that it is unlikely that this optimal point falls at either extreme. Therefore, industrial inventories should benefit from the judicious inclusion of CFI features. It is not widely acknowledged that this is possible because the above scales are continuous, and the periodic inventories would not have to conform exactly to strict CFI specifications. Hence, we arrive at PSEUDO-CFI, defined as the family of sample designs that lie between periodic inventories and continuous forest inventory, obtained by formally relaxing the requirements of CFI, hopefully in a manner that nets the greatest decrease in cost for the least increase in variance.

Although the term "continuous forest inventory" or "CFI" has been applied to many different designs, most have a common set of elements:

1. Permanent sample points are used,
2. Regularly scheduled remeasurements are made,
3. Individual trees are stem-mapped, and
4. A reconciliation of sample trees with those existing at the previous measurement is made.

All these elements could be relaxed in different ways, although some would contribute to such an increase in variance that "borrowing" from CFI would be of little worth.

There are many ways in which the first requirement, of using permanent sample points, might be relaxed. This allows a range of choices with respect to cost per unit variance reduction. (Of course, the result of eliminating it completely would be a periodic inventory). It is expected that most useful designs will be somehow related to a permanent list of points. For instance, we could have a permanent list of sample points but not require exact field relocation of all (or any) of the points. Because the cost of approximate relocation of point center is low, while the cost for exact (or very small tolerance) point center location can be relatively high, it could be advantageous to formulate a sample design and the corresponding estimation system for particular levels of relaxation of the relocation requirement. One idea would be to limit the search time for point center once the general location has been reached. When the limit is exceeded, the data could be taken at the "best guess" point location. Alternatively a requirement might be set that each point must be located within +/- a maximum distance, such as that achievable through a geographic positioning system. A workable plan would be to require that x percent of the points be relocated to within a very small tolerance, while the remaining (100-x) percent of the points be relocated to within a larger tolerance.

As with the point relocation requirement, there are many ways in which the regular timing requirement could be relaxed. Feast or famine cycles could be allowed in timing of measurement periods. This would be attractive to landowners that experience fluctuations in business activity. These types of businesses would benefit from being able to collect most of the data while adequate revenues are being generated. Related ideas include allowing different interval lengths between measurements based on a priority classification, or devising a temporally defined subsampling scheme.

Another aspect of CFI, which could be relaxed to various degrees, is the rigor with which sample trees are identified for future relocation. Often trees are both physically marked and their locations are entered into the database. At each successive visit, an extensive reconciliation of the previous measurement's sample trees is undertaken to allow the fitting of individual tree and spatially dependent models. One way to relax this requirement is to define a minimum diameter, below which the trees are not mapped or tagged. Not individually identifying any sample trees is another option. Or tagging could be chosen over stem-mapping because physically marking the trees can be relatively inexpensive while stem-mapping with all its implications, such as having the previous inventory's individual tree data available to the field crew, is relatively expensive.

Relaxing the reconciliation requirement could take the form of requiring only the reconciliation of a subpopulation of the trees, such as the largest or most valuable trees, or of only a subsample of trees. If only a subpopulation of trees is reconciled, then individual tree models could not be calibrated for the remaining population. Subsampling from the entire population is the more satisfying method of relaxing this requirement. Of course, any number of tree subsampling schemes could be devised for permanent identification.

Without much effort, we could use the relaxation of both tree and point relocation and have, as a result, a system that has CFI/SPR (sampling with partial replacement) design qualities. For instance, we could require exact relocation on a subset of points and approximate relocation on the remaining points at short time intervals, while requiring exact relocation of all points at a longer time interval. The degree to which trees are remeasured could be allowed to depend on the point relocation criterion. We would thereby achieve most of the benefit of a CFI design with respect to variance of the growth estimate without incurring the high cost of exactly relocating each point at each visit.

This plan begs the question: how close would we have to get on the approximately relocated plots? The answer to this question is case specific, because it depends on the proportion of exactly relocated plots, the time intervals between measurements, and the growth rates of the populations of interest. Note that you would even expect a growth estimator formed on the less than perfectly remeasured part of the sample (i.e., those trees not included at both measurements) to have a smaller variance than one formed from two spatially independent samples because neighboring trees often share many of the same characteristics. In addition, through time, if point relocation errors were randomly distributed, you would expect trees that are missed at one remeasurement to have an equal probability of being measured in a subsequent time period as trees that were not missed. However, point relocation errors may not be randomly distributed. Further simulation study is needed in this area.

ANALYSIS

Data from a Pseudo-CFI design might be used in several ways. During the first few measurement periods, the data would be used only to adjust the resulting updated estimates. After the design has been in place for a number of measurement periods, the data could be used to directly calibrate the growth and yield equations.

A long-range analytical plan would be to treat existing growth and yield estimators as a prior distribution, collect data on multidimensionally defined subpopulations, and use it to calibrate the original estimators into final or posterior growth and yield estimators for appropriate segments of the subpopulations. This approach would help to correct the erroneous assumption in traditional growth and yield modeling that forest growth is stationary in the statistical sense with respect to time and space.

Mixed Estimation

Because we have information from a wide time interval and want it to give us as accurate a picture as possible of the current state, we should formally account for the fact that estimates derived from more recent data will have a lower variance than those derived from older data. Theil (197 1) provides a solution known as mixed estimation, which was discussed in Van Deusen (1989, 1996). The latter work is followed directly here.

For simplicity, assume that we have a large sample, a survey cycle of T time periods, and that we are interested in sample plots that have not been harvested during the cycle. A simple model for our sample data at time $t=1,\ldots,T$ is:

$$\bar{y}_t = \mu_t + \bar{e}_t$$

where $\bar{e}_t$ is an error term with a mean of 0 and a variance of σ_t^2/n_t, which we would estimate by the usual sample estimator. Now let $\hat{y}_t$ represent the prediction obtained from the growth model at time t and $\hat{g}_t = \hat{y}_t - \hat{y}_{t-1}$, where $t = 2,\ldots,T$. A simple model for the time series might then be represented as:

$$\hat{g}_t = \mu_t - \mu_{t-1} + v_t$$

where v_t is an error term of zero mean and λ_t^2/m_t variance.

Collect the $\bar{y}_t$'s into the vector $\bar{\mathbf{Y}} = [\bar{y}_1,\ldots,\bar{y}_T]'$, the $\hat{g}_2$'s into the vector $\hat{\mathbf{G}} = [\hat{g}_2,\ldots,\hat{g}_T]'$, the means into the vector $\mathbf{U} = [\mu_1,\ldots,\mu_T]'$, and the error terms into the vectors $\mathbf{e} = [\bar{e}_1,\ldots,\bar{e}_T]'$ and $\mathbf{v} = [v_2,\ldots,v_T]'$.

The matrix representation of the sample data model is then $\overline{\mathbf{Y}} = \mathbf{U} + \mathbf{e}$.

With the definition of two design matrices:

$\mathbf{D_1}$, a TxT matrix in which the diagonal elements are 1 and the off-diagonals are 0 and

$$\mathbf{D_2} = \begin{bmatrix} -1 & 1 & 0 & . & . & . & 0 \\ 0 & -1 & 1 & . & . & . & 0 \\ . & & & . & & & . \\ 0 & 0 & \ldots & & -1 & & 1 \end{bmatrix}$$, a T-l by T matrix,

we can represent the time series as $\hat{\mathbf{G}} = \mathbf{D_2}\mathbf{U} + \mathrm{v}$. Furthermore, we can write the equations in the combined form as:

$$\begin{bmatrix} \overline{\mathbf{Y}} \\ \hat{\mathbf{G}} \end{bmatrix} = \begin{bmatrix} \mathbf{D_1} \\ \mathbf{D_2} \end{bmatrix} \mathbf{U} + \begin{bmatrix} \mathbf{e} \\ \mathbf{v} \end{bmatrix}$$

Define:

$$\Sigma_1 = \begin{bmatrix} \sigma_1^2/n_1 & \sigma_{12}/n_{12} & \cdots & \sigma_{1T}/n_{1T} \\ \sigma_{21}/n_{21} & \sigma_2^2/n_2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{T1}/n_{T1} & \cdots & \cdots & \sigma_T^2/n_T \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} \lambda_2^2/m_2 & \lambda_{23}/m_{23} & \cdots & \lambda_{2T}/m_{2T} \\ \lambda_{32}/m_{32} & \lambda_3^2/m_3 & \cdots & \\ \vdots & \cdots & \ddots & \\ \lambda_{T2}/m_{T2} & & \cdots & \lambda_T^2/m_T \end{bmatrix}$$

With independent periodic inventories, the off-diagonal elements in the above two matrices can be assumed to be zero. Although convenient, this assumption would not be valid or desirable in the present context.

The mixed estimator would be:

$$\hat{\mathbf{U}} = \left(\mathbf{D_1}'\sum{}_1^{-1}\mathbf{D_1} + \mathbf{D_2}'\sum{}_2^{-1}\mathbf{D_2}\right)^{-1}$$

$$\left(\mathbf{D_1}'\sum{}_1^{-1}\overline{\mathbf{Y}} + \mathbf{D_2}'\sum{}_2^{-1}\hat{\mathbf{G}}\right)$$

with a covariance matrix:

$$\mathbf{V}\left(\hat{\mathbf{U}}\right) = \left(\mathbf{D_1}'\sum{}_1^{-1}\mathbf{D_1} + \mathbf{D_2}'\sum{}_2^{-1}\mathbf{D_2}\right)^{-1}$$

EXAMPLE APPLICATION – NORTHEASTERN UNITED STATES

The dynamics of forest populations in the northeastern spruce-fir forest are due largely to a spruce-budworm outbreak in the 1980’s. Spruce-budworm effects were not a factor considered in the growth and yield functions available for updating the inventory, and the current populations are quite different from any of those upon which the existing models were built. Therefore, a Pseudo-CFI sample design was overlaid on an existing systematic design to continuously monitor and possibly calibrate the dynamic inventory to the changing populations.

The design starts with a permanent randomly placed grid of points (20 by 5 chain spacing) laid over an entire region. The grid points are sampled in an extensive “base” inventory every 15 years, and subsampled every 3 years. We give details of the procedure below.

Sampling Procedure:

1. Photographs of the inventory area will be interpreted to determine the stratum and priority class of each stand every 15 years. The stratum classification will be continuously updated as growth, silvicultural practices, or harvesting actions change a polygon’s stratum membership. Each stratum will be classified into one of three spatial sampling intensities, labeled high, medium, and low priority.

2. A permanent base grid of points (5 by 20 chains) will be randomly placed over the region. All points falling outside of the areas of interest are ignored, and the remaining points are numbered from 1 to N. The grid points actually used in the full field sample will be selected every 15 years from these grid points weighted by priority class. Every point. in every line will be measured in high priority strata, every even-numbered point in every line will be measured in medium priority strata, and every even-numbered point in every fifth line will be measured in low priority strata. (The starting line will be chosen at random from the first five lines.) Since the grid is permanent, most of these points will be relocated as closely as possible and remeasured at the next full inventory in 15 years; only adjustments due to priority class changes will be made.

3. Every 3 years a systematic subsample of 15 percent of the sample points will be remeasured. During the

initial inventory, the sample points will be numbered from 1 to n beginning at the first sample point of line 1. Then a random number between 1 and 100 will be chosen as the starting point. This and the next 14 contiguous sample points will be included. The next 85 sample points will be skipped. This pattern will be repeated until all sample points in the grid have been worked through. The subsample points will be located with a GPS and monumented. The 15-point cluster was chosen because this is expected to be the number of points on which data can be collected in 1 day (at least in fully stocked stands). All point-level and tree-level measurements will be made. "In" trees greater than 4.5 inches dbh will be measured in a clockwise direction from due north and tagged with an identification number.

CONCLUSION

Pseudo-CFI is a family of sample designs that exploit a particular section of the continuum of potential degree of control in investigations of forest growth and yield. This family of Pseudo-CFI sample designs is not intended to replace controlled experiments for the development of growth and yield equations. It is merely intended as a monitoring mechanism to facilitate the identification of badly behaving populations and provide a reasonable set of data that can be used to calibrate model estimates for these populations. It is arguable that all forest populations are constantly moving away from those upon which their respective growth and yield models were built. What determines the necessity for model calibration is the rate at which the population is moving in relation to the length of time that has elapsed since model development. Because this can not be known in the absence of a monitoring effort, these ideas can and should be applied to all dynamic inventory systems, not just to industrial systems. The reason industrial inventory systems have been stressed here is that these systems, in general, have a much greater investment per acre and have been using dynamic features longer than public sector inventories.

ACKNOWLEDGMENTS

Phil Malerba, Technical Services Manager, Northeast Region, Champion International Corporation, provided many insights that helped to identify the need for Pseudo-CFI. In addition, he provided much of the detail used to convert the theory into an operationally efficient design. The following people reviewed this paper: Dr. William Bragg, Champion International Corporation, Jacksonville, FL, USA, and Dr. Gregory A. Reams, Southern Research Station, USDA Forest Service, Asheville, NC, USA.

LITERATURE CITED

Basu, D. 1980. Rejoinder to comments on randomization analysis of experimental data: The Fisher randomization test, by D. Basu. Journal of the American Statistical Association. 75: 575-595.

Cochran, W.G. 1965. The planning of observational studies of human populations. Journal of the Royal Statistical Society A. 128: 234-265, with discussion.

Cochran, W.G.; Rubin, D.B. 1974. Controlling bias in observational studies: a review. Sankhyä A 35: 417-446.

Green, E.G.; Roesch, F.A., Jr. 1993. Statistical analysis of observational studies with emphasis on forest response studies. In: Rennolls, K.; Gertner, G., eds. The optimal design of forest experiments and forest surveys, Proceedings of a IUFRO S4.11 conference; 1991 September 10-14, London UK: University of Greenwich: 157- 166.

Rubin, D.B. 1973a. Matching to remove bias in observational studies. Biometrics. 29: 159- 183.

Rubin, D.B. 1973b. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics. 29: 185-203.

Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 66: 688-70 1.

Rubin, D.B. 1976a. Multivariate matching methods that are equal percent bias reducing, I: some examples. Biometrics. 32: 108: 120.

Rubin, D.B. 1976b. Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. Biometrics. 32: 121-132.

Rubin, D.B. 1978. Bayesian inference for causal effects: The role of randomization. The Annals of Statistics. 6: 34:58.

Rubin, D.B. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association. 74: 3 18-328.

Rubin, D.B. 1980. Discussion of randomization analysis of experimental data: The Fisher randomization test, by D. Basu. Journal of the American Statistical Association. 75: 59 l-593.

Rubin, D.B. 1986. Which ifs have causal answers. Comments on statistics and causal inference by P.W. Holland. Journal of the American Statistical Association. 81: 961-962.

Theil, H. 197 1. Principles of econometrics. New York: John Wiley & Sons.

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

Van Deusen, P.C. 1996. Incorporating predictions into an annual forest inventory. Canadian Journal of Forest Research. 26: 1709- 17 13.

Wickramarante, P.J.; Holford, T.R. 1987. Confounding in epidemiologic studies: the adequacy of the control group as a measure of confounding. Biometrics. 43: 752-765.

Wold, H. 1956. Causal inference from observational data. Journal of the Royal Statistical Society A. 119: 28-60, with discussion.