

Reprinted from *Forest Science*, Vol. 39, No. 1, February 1993.

Control Variate Estimators of Survivor Growth from Point Samples

F.A. ROESCH, JR.
P.C. VAN DEUSEN

ABSTRACT. Two estimators of the control variate type for survivor growth from remeasured point samples are proposed and compared with more familiar estimators. The large reductions in variance, observed in many cases for estimators constructed with control variates, are also realized in this application. A simulation study yielded consistent reductions in variance which were often 50% and occasionally over 70%. *FOR. SCI.* 39(1):66-77.

ADDITIONAL KEY WORDS. Point sampling, simulation, forest growth, forest inventory, control variates.

THE GROWTH OF FORESTS IS OF PARAMOUNT IMPORTANCE in forestry, and quite often foresters are interested in the specific components of growth between two points in time. These components of growth are ingrowth (I), survivor growth (S), mortality (M), and cut (C), and computing these components from remeasured point samples has been the subject of several recent articles (i.e., Martin 1982, Van Deusen et al. 1986, and Roesch et al. 1989). Quite often these articles are concerned with achieving compatible estimates of these components. Compatible estimators are those that are constrained to maintain the equality (Meyer 1953):

$$\hat{V}_2 - \hat{V}_1 = \hat{S} + \hat{I} - \hat{M} - \hat{C},$$

where

\hat{V}_i = estimator of the value of interest at time i ,
 \hat{S} = estimator of survivor growth,
 \hat{I} = estimator of ingrowth,
 \hat{M} = estimator of mortality, and
 \hat{C} = estimator of cut.

The advantages and disadvantages of compatible estimators for the components of growth have been discussed at length in the literature. A very complete discussion can be found in Flewelling (1981), for example. Other discussions appear in Martin (1982), Van Deusen et al. (1986), and Roesch et al. (1989). While compatibility is a highly desirable condition when estimates of all of the components of change in a value, along with estimates of the value at each time are desired, it is not an important consideration when only one or a few of these estimates are desired. Even when all of these estimates are to be made, many would argue that minimum mean squared error of each estimator, for example, is a more important criterion than compatibility.

When compatibility is not the major concern, it is of greater interest to obtain the best estimator of each of the components. That is why, in this paper, we will not be concerned with forcing compatibility and will concentrate on a method of reducing the variance of the individual estimators. We will look specifically at estimators of survivor growth after reviewing a general variance reduction technique found in Rubinstein (1981). A simulation comparison is made of the basal area survivor growth estimators discussed in Grosenbaugh (1958), Van Deusen et al. (1986), and Roesch et al. (1989) with estimators using control variates.

ESTIMATOR CONSTRUCTION WITH CONTROL VARIATES

Rubinstein (1981) shows how a control variate (C) could be used to reduce the variance of an estimator. The control variate is correlated with and used in conjunction with an estimator (say \hat{E}_1) of a parameter θ to produce another estimator (\hat{E}_2) of equal bias and smaller variance. If the mean of the control variate (μ_c) is known, then for any β

$$\hat{E}_2 = \hat{E}_1 - \beta(C - \mu_c) \quad (1)$$

is an estimator of θ with a bias equal to that of \hat{E}_1 .

Since

$$\text{var}[\hat{E}_2] = \text{var}[\hat{E}_1] - 2\beta\text{cov}[\hat{E}_1, C] + \beta^2\text{var}[C], \quad (2)$$

the variance of \hat{E}_2 will be lower than that of \hat{E}_1 if

$$2\beta\text{cov}[\hat{E}_1, C] > \beta^2\text{var}[C]. \quad (3)$$

When β is set equal to

$$\beta^* = \text{cov}[\hat{E}_1, C]/\text{var}[C], \quad (4)$$

the variance of \hat{E}_2 will be minimized and equal to

$$(1 - \rho^2)\text{var}[\hat{E}_1], \quad (5)$$

where ρ is the correlation coefficient between \hat{E}_1 and C. In most cases β^* will be estimated from the sample as outlined in Appendix 1.

SURVIVOR GROWTH ESTIMATORS

The notation in this section will follow that of Martin (1982). First, recall that a per acre estimate from a point sample is obtained by summing the ratio v/ba (where v and ba are the value of interest and basal area, respectively) over all trees on the point and multiplying by the basal area factor.

For our purposes, survivor growth (S) over a specific time interval occurs on the population of trees which were in some specific merchantability range during the entire interval. Survivor trees are sampled in two ways on remeasured points. The n sample is measured only at time 2 (i.e., these trees were "out" at time 1) while the s sample is measured at both times 1 and 2.

Grosenbaugh (1958) advocated the use of only the s sample for the survivor growth estimator:

$$\hat{S} = s_2' - s_1 \quad (6)$$

since the n sample is confounded with the sample of the population of ingrowth

trees which was only collected (measured) at time 2 (we will call this sample of ingrowth trees the o sample). In Equation (6):

s_2' = estimate obtained using the *final* values of v and the initial values of ba for the survivor trees in the s sample; and

s_1 = estimate obtained using the *initial* values of v and the initial values of ba for trees in the s sample.

In addition we define:

s_2 = estimate obtained using the *final* values of v and the *final* values of ba for trees in the s sample;

s_1' = estimate obtained using the *initial* values of v and the *final* values of ba for trees in the s sample;

n_2 = estimate obtained using the *final* values of v and the *final* values of ba for survivor trees measured only the second time (the n sample);

n_1' = estimate obtained using the estimated *initial* values of v and the *final* values of ba for trees in the n sample;

Following Martin (1982) we use a subscript to refer to the time of measurement (1 or 2) of v and a prime (') to indicate that ba in the above ratio is from the time other than that subscripted.

\hat{S} is well defined and easy to use, but it does not fully utilize the information available from a remeasured point sample. To use any more information in a survivor growth estimator, an investigator must assume that (s)he can adequately distinguish between the two populations (ingrowth trees and survivor trees) in order to separate the o and the n samples.

Van Deusen et al. (1986), recognizing the information potential of the n sample, proposed the following estimator for S :

$$\tilde{S} = s_2 - s_1 + n_2. \quad (7)$$

Another estimator,

$$S^{**} = s_2 - s_1' + n_2 - n_1', \quad (8)$$

was defended by Roesch et al. (1989) as using only the growth of the sampled survivor trees (that measured on s and estimated on n).

To facilitate our construction of the control-variate type estimators, we recall that Van Deusen et al. (1986) showed that \tilde{S} differed from \hat{S} by an error term:

$$\tilde{S} = \hat{S} + \epsilon, \quad (9)$$

where

$$\epsilon = s_2 - s_2' + n_2, \quad (10)$$

and that ϵ has an expected value of zero if predictions of the time 1 values of trees measured only at time 2 are unbiased enough to separate ingrowth and survivor trees. In addition, Roesch et al. (1989) showed that ϵ can be decomposed into two components, each having an expected value of zero given the above condition:

$$\epsilon = \epsilon_1 + \epsilon_G, \quad (11)$$

where

$$\epsilon_1 = \{s_1' - s_1 + n_1'\}, \text{ and} \quad (12)$$

$$\epsilon_G = \{(s_2 - s_1') - (s_2' - s_1) + (n_2 - n_1')\}. \quad (13)$$

They then showed that S^{**} can be written as:

$$S^{**} = \hat{S} + \epsilon_G. \quad (14)$$

The error terms above could also be viewed as control variates, in the spirit of Rubinstein (1981). This leads to two new estimators for survivor growth:

$$\tilde{S}_o = \hat{S} - \beta^{*1}(\epsilon - \mu_\epsilon), \quad (15)$$

and

$$s_o^{**} = \hat{S} - \beta^{*2}(\epsilon_G - \mu_{\epsilon_G}), \quad (16)$$

where the β^{*j} s are estimated using Equation (4), and μ_ϵ and μ_{ϵ_G} are the means of ϵ and \bullet G, respectively (assumed in the sequel to be equal to zero). It can be shown that each of the two control variate estimators above is unbiased.

We will conduct a simulation study to demonstrate that an appreciable reduction in variance and mean square error could be achieved by either of the control variate estimators in Equations (15) and (16). The study will show that the control variate estimators, while not always "best" by every criterion, appear to always be as good as or better than S , S^{**} , and \tilde{S} in terms of squared error loss under widely varying conditions. We would suspect this possibility since the information base is smaller for S than for the other four estimators and the control-variate estimators use this information more efficiently than their counterparts.

METHODS

SIMULATION DESCRIPTION

The data came from a mutual competition study¹ of loblolly pine (*Pinus taeda* L.) planted in February 1958. In March 1962, four ¼ ac stem-mapped plots for each of five post-thinning target densities (treatments) were established. The target densities of treatments 1-5 were 1000, 600, 300, 200, and 100 trees/ac, respectively. Each tree was measured 8 times between 1962 and 1981, inclusive (1962, 1963, 1965, 1966, 1967, 1970, 1977, and 1981).

In the simulation, we used four treatment/time interval combinations which were known from previous simulations (Roesch 1989) to represent the four different relationships of variance and bias observed over all the possible treatment/time interval combinations between the basal area survivor growth estimators S , \tilde{S} , and S^{**} . In cases 1 and 2 the variances of \tilde{S} and \hat{S} are about equal. In case 1 the mean square error (MSE) of these two estimators are also equal, while in case 2 the MSE of \tilde{S} is higher than that of \hat{S} . In cases 3 and 4, one of the basic estimators outperforms the other in terms of variance and MSE. Case 3 favors S

¹ Hill Farm Research Station of the Louisiana Agricultural Experiment Station, Louisiana State University Agricultural Center in Homer, LA.

over \hat{S} and the opposite is true in case 4. We are doing this under the assumption that the same factors which result in differences between \hat{S} , \hat{S} , and S^{**} will also expose potential differences between these estimators and the control-variate estimators. We used treatment 5 of the 11-yr growth interval from 1970 to 1981, treatment 5 from the 10-yr growth interval from 1967 to 1977, treatment 2 of the 11-yr interval from 1970-1981, and treatment 4 of the 4-yr interval from 1977-1981 to represent cases 1 to 4 respectively. We used a 9 in. merchantability limit in all four cases, corresponding to trees which qualify as softwood sawlog trees.

For each case we simulated a square 900 ac forest from the four $\frac{1}{4}$ ac plots randomly arranged in a line 900 times until the square was filled from the top left to the bottom right (i.e., each arrangement forms a rectangular acre which is 104.355 ft wide and 417.421 ft long). Then we repeatedly sampled the 900 ac forest by taking 50 sets of 50 randomly located points for every basal area factor (BAF) divisible by 5 from 5 to 50 ft²/ac inclusive. Edge bias was eliminated by flipping the edge plots out into the border around the forest and also including any "in" trees from the "outside" plots while only allowing the sample points to fall within the 900 acre forest. For each sample we calculated the mean square error, variance and bias for the estimates of survivor basal area growth obtained from each of the estimators above and then plotted these values against BAF for each case. At time 2 in each case we predicted the previous time's basal areas of the "new" trees in the sample by using a weighted linear regression of time 1 basal area on time 2 basal area for the "old" trees. Because "s" is a sample selected proportional to ba_1 , while n is a sample selected proportional to $(ba_2 - ba_1)$, a weight must be applied to the regression for the "old" sample trees to obtain unbiased predictions of time 1 basal areas of the "new" sample trees. This weight is derived in Appendix 2. Granted, a linear model may not always be the best model in a particular case. However, a thorough model evaluation in each case would be extremely time consuming and not germane to this discussion. In fact, all of the estimators of survivor growth with the exception of S are more or less model dependent, and the cases we are presenting are at least partially artifacts of the model selected. Therefore, we are, in part, investigating how robust each estimator is to the model selection process.

RESULTS

As we mentioned above, cases 1 to 4 represent the four basic patterns observed at the conclusion of the preliminary sampling simulations reported by Roesch (1989). Figures 1 through 4 contain the results for cases 1 to 4, respectively. Figures 1 and 2 represent the cases in which the variances of S and \hat{S} are about equal. Figure 1 shows the mean square error (MSE) of these two estimators to also be equal while Figure 2 shows the MSE of \hat{S} to be higher than that of \hat{S} . The difference is due to bias in \hat{S} resulting from biased predictions of the time 1 basal areas of trees in the new sample (n). This bias occurs in some cases in spite of the weighting described above and is due to the fact that the growth curve is at least somewhat nonlinear. In Figure 1, we see that the two control variate estimators have about the same variance as each other and S^{**} , while all of the estimators appear relatively unbiased. The two basic estimators (\hat{S} and \hat{S}) are performing quite well for the time interval, growth rate and BAFs considered, but

11-year growth interval (1970-1981), Treatment 5
Sawlog Growth

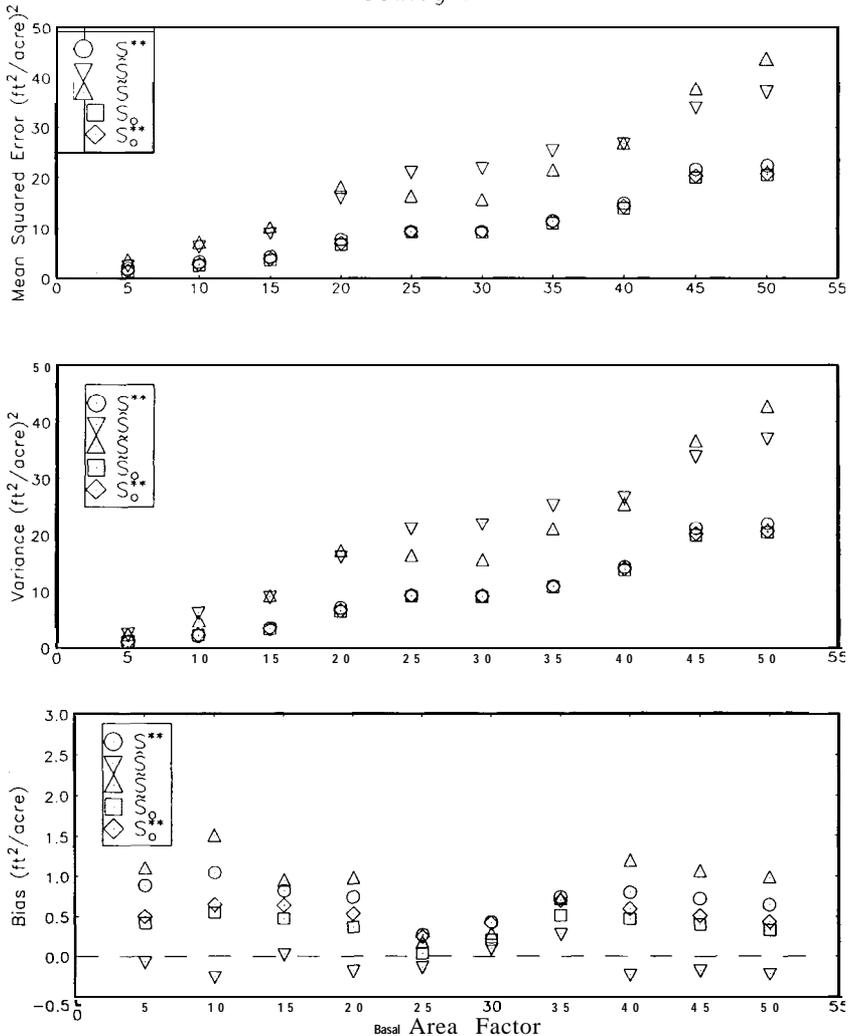


FIGURE 1. The mean squared error, variance, and bias of the estimators of sawlog-sized basal area survivor growth versus BAF. The actual basal area growth was 31.18 ft²/ac.

the control variate estimators are performing even better with reductions in variance and MSE of over 50% in many cases.

Figure 2, in which the growth interval (1967-1977) is greater than the second half of the elapsed life of the plantation (1958-1977), tells quite a different story, as the previous basal areas become harder to predict. Gains using the control variate estimators over (\hat{S}) in terms of variance reduction are similar to but slightly less than those found in Figure 1. The performances of \hat{S}_o and \hat{S}_o^{**} are comparable to those observed in Figure 1 overall despite the fact that prediction bias is evident for \hat{S} .

Figures 3 and 4 represent the cases in which one of the basic estimators

1 O-year growth interval (1967–1977), Treatment 5
Sawlog Growth

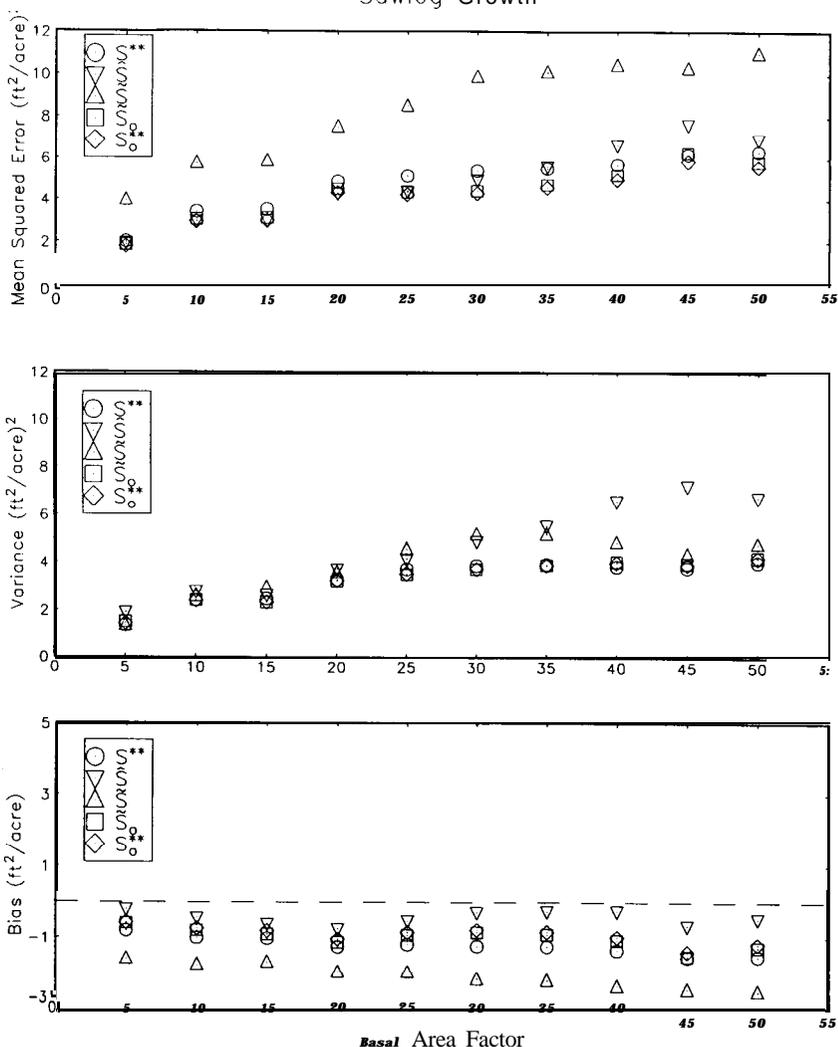


FIGURE 2. The mean squared error, variance, and bias of the estimators of sawlog-sized basal area survivor growth versus BAF. The actual basal area growth was 6.82 ft²/ac.

outperforms the other in terms of variance and MSE. Figure 3 favors \hat{S} over \hat{S}_x at least at the larger BAFs, and the opposite is true in Figure 4. In Figure 3, the two control variate estimators, which are constructed from \hat{S} , perform very well in terms of MSE when compared to \hat{S}_x with reductions of 50% common. Results like these could occur when data are sparse; in this instance not many of the trees had reached the sawlog category by 1970.

Figure 4 represents cases in which the variance and MSE of \hat{S} are higher than those of \hat{S}_x , although the prediction bias problem does not seem to be a great concern. When the time interval is relatively short, not allowing many “new” trees to enter the sample, and a fair number of trees were measured at time 1, results

11 -year growth interval (1970–1981), Treatment 2
Sawlog Growth

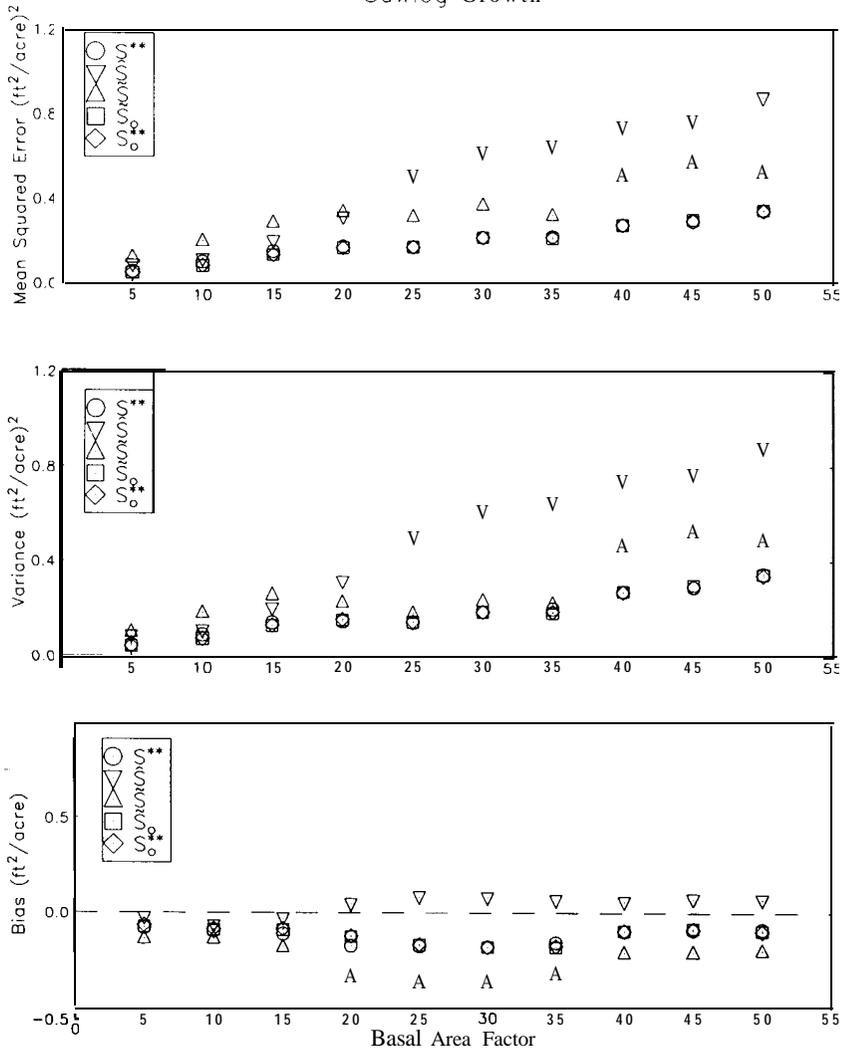


FIGURE 3. The mean squared error, variance, and bias of the estimators of sawlog-sized basal area survivor growth versus BAF. The actual basal area growth was $0.66 ft^2/ac$.

such as these will occur. In these cases the variances and MSEs of \tilde{S}_o , S_o^{**} , and \tilde{S}_o^{**} are just a little lower than those of \hat{S} .

CONCLUSIONS

In this manuscript we have presented two control-variate type estimators, \tilde{S}_o and S_o^{**} , which could be viewed as the “optimized” versions of \hat{S} and S^{**} . A simulation demonstrated the properties developed in theory; i.e., that \tilde{S}_o and S_o^{**} must be better than or equal to \hat{S} in terms of variance (and therefore mean square error). This simulation presents strong evidence in support of the overall useful-

4-year growth interval (1977–1981), Treatment 4
Sawlog Growth

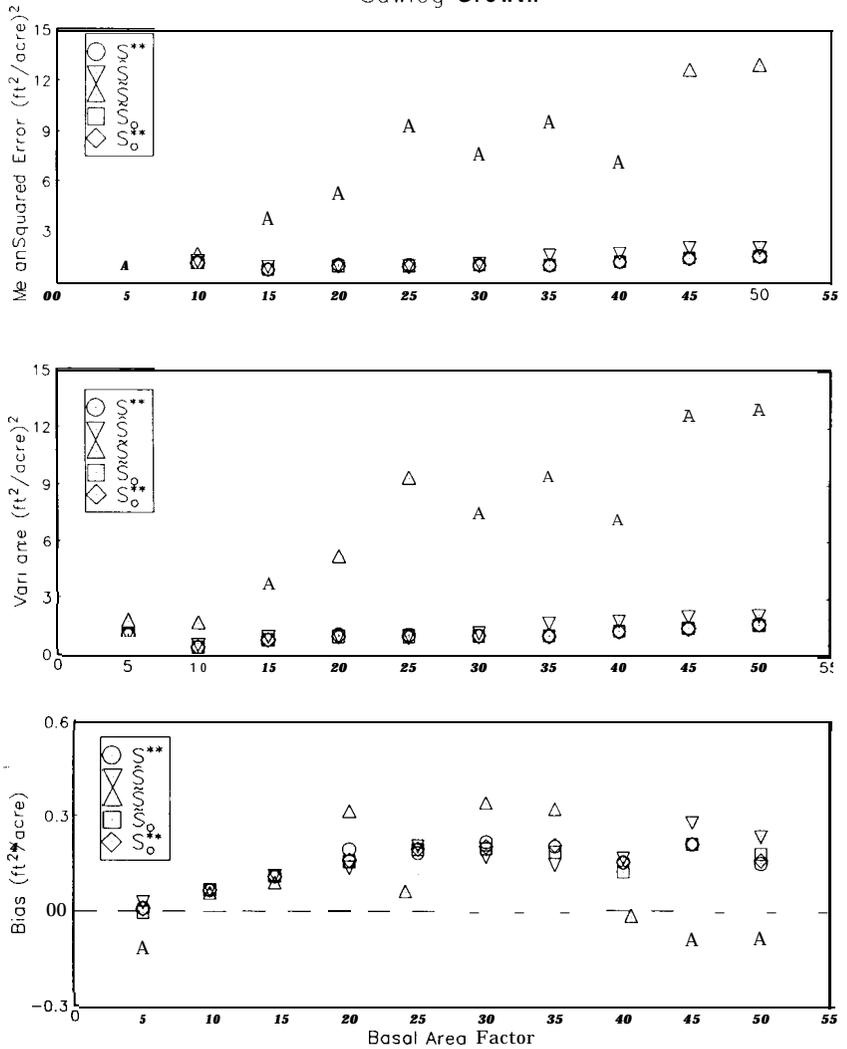


FIGURE 4. The mean squared error, variance, and bias of the estimators of sawlog-sized basal area survivor growth versus BAF. The actual basal area growth was 14.00 ft²/ac.

ness of these control variate estimators. Substantial decreases in variance have been realized, without the control variate estimators being as susceptible to prediction bias as are both \hat{S} and \hat{S}^{**} . We remind the reader that our choice of a linear model had, in this simulation, some contribution to the bias of \hat{S} and to a lesser extent \hat{S}^{**} . Our intent is not to make absolute comparisons but to show that even when conditions make previous predictions difficult, the control variate estimators still work well. Results like those of case 1 could be expected quite commonly. The other three cases might be considered pathological in some sense, although they probably occur often enough in real inventories to be of concern.

There doesn't seem to be a clear choice between the control variate estimators

in terms of variance or bias, although \hat{S}_o might be favored because it is simpler to implement than S_o^{**} . S^{**} also behaved well in terms of variance and MSE; very often as well as S_o and S_o^{**} . However, the potential for prediction bias to occur would seem to negate the advantages of S^{**} when compatibility is not an issue.

LITERATURE CITED

- FLEWELLING, J.W. 1981. Compatible estimates of basal area and basal area growth from re-measured point samples. *For. Sci.* 27(1):191-203.
- GROSENBAUGH, L.R. 1958. Point-sampling and line-sampling: Probability theory, geometric implications, synthesis. USDA For. Serv. Occas. Pap. 160. 34 p.
- MARTIN, G.L. 1982. A method of estimating ingrowth on permanent horizontal sample points. *For. Sci.* 28(1):110-114.
- MEYER, H.A. 1953. Forest mensuration. Penn Valley Publishers, State College, PA. 357 p.
- ROESCH, F.A. JR. 1989. Growth from variable-radius plots, solutions revisited. Unpubl. pap. presented at the Assoc. of Southern University Biometricians Annu. Meet.
- ROESCH, F.A., JR., E.J. GREEN, and C.T. SCOTT. 1989. New compatible estimators for survivor growth and ingrowth from re-measured horizontal point samples. *For. Sci.* 35(2):281-293.
- RUBINSTEIN, R.Y. 1981. Simulation and the Monte Carlo method. Wiley, New York, 278 p.
- VAN DEUSEN, P.C., T.R. DELL, and C.E. THOMAS. 1986. Volume growth estimation from permanent horizontal points. *For. Sci.* 32(2):415-422.

Copyright 1993 by the Society of American Foresters
 Manuscript received July 8, 1991

AUTHORS AND ACKNOWLEDGMENTS

Francis A. Roesch, Jr., and Paul C. Van Deusen are Mathematical Statisticians, Institute for Quantitative Studies, Southern Forest Experiment Station, USDA Forest Service, 701 Loyola Avenue, New Orleans, LA 70113. This work was supported by the USDA Forest Service Health Monitoring Program.

DISCUSSION

The method to construct a generic control variate estimator (\hat{E}_2) of basal area survivor growth from another estimator (\hat{E}_1) of survivor growth and a control variate (C) of known mean (μ_c) is given below. In the paper we use control variates with known means of zero. The covariance/variance ratio must be estimated from the data. To do this, calculate both \hat{E}_1 and C at each point and determine the sample estimates of both the covariance of \hat{E}_1 and C and the variance of C :

$$\text{cov}(\hat{E}_1, C) = \sum_{i=1}^m ((\hat{E}_{1i} - \bar{\hat{E}}_1) * (C_i - \bar{C})) / m - 1 \quad (1.1)$$

and

$$\text{var}(C) = \sum_{i=1}^m ((C_i - \bar{C}) * (C_i - \bar{C})) / m - 1; \quad (1.2)$$

where

m = the number of sample points,
 \hat{E}_{1i} = the value of \hat{E}_1 at point i ,

$$\bar{\hat{E}}_1 = \sum_{i=1}^m (\hat{E}_{1i}/m),$$

C_i = the value of C at point i , and

$$\bar{C} = \sum_{i=1}^m (C_i/m).$$

From 1.1 and 1.2 above, an estimate of β^* is obtained:

$$\beta^* = \text{cov}(\hat{E}_1, C) / \text{var}(C). \quad (1.3)$$

Finally, the unbiased control variate estimator is attained:

$$\hat{E}_2 = \hat{E}_1 - \beta^*(\bar{C}). \quad (1.4)$$

APPENDIX 2

To predict the previous basal areas of the trees measured only at time 2 (the “new” sample) using the regression from the trees measured both times (the “old” sample), we assume that trees do not shrink during the growth interval. Under this assumption the old sample is weighted proportional to basal area at time 1 while the new sample is weighted proportional to basal area growth. Therefore, unbiased predictions of the time 1 values for the new sample require that weighted regression be used. This appendix shows how to weight the regression of basal area at time 1 (b_1) on basal area at time 2 (b_2) for the “old” sample in order to obtain an unbiased predictor of b_1 , for trees in the “new” sample.

The regression of interest for the “new” sample is simply the expected value of b_1 , given b_2 , under the sample domain. Given the large sample of all trees of a particular merchantability class measured at time 2, the probability that a particular tree was not measured at time 1 (i.e., it is in the “new” sample) is equal to $(b_2 - b_1)/b_2$. So the expected value of time 1 basal area for trees of a particular b_2 in the “new” sample can be determined using the fundamental properties of conditional expectation:

$$\begin{aligned} E_{\text{new}}[b_1|b_2] &= E[[b_1|b_2]|\text{“new”}] \\ &= \frac{E[[b_1|b_2], \text{“new”}]}{E[\text{“new”}]} \end{aligned} \quad (2.1)$$

Therefore, by the definition of expected value

$$\begin{aligned}
E_{new}[b_1|b_2] &= \frac{\int_0^\infty b_1 \left(\frac{(b_2 - b_1)}{b_2} \right) f(b_1|b_2) db_1}{\int_0^\infty \left(\frac{(b_2 - b_1)}{b_2} \right) f(b_1|b_2) db_1} \\
&= \frac{b_2 E[b_1|b_2] - E[b_1^2|b_2]}{b_2 - E[b_1|b_2]}
\end{aligned} \tag{2.2}$$

after cancellation of the b_2 terms in the denominator of each integral. Note that E_i indicates the expected value in sample domain i and if E is unsubscripted, the expectation is over the entire population.

Likewise the same expectation for the “old” sample:

$$\begin{aligned}
E_{old}[b_1|b_2] &= \frac{\int_0^\infty b_1 \left(\frac{b_1}{b_2} \right) f(b_1|b_2) db_1}{\int_0^\infty \left(\frac{b_1}{b_2} \right) f(b_1|b_2) db_1} \\
&= \frac{E[b_1^2|b_2]}{E[b_1|b_2]}.
\end{aligned} \tag{2.3}$$

Equations (2) and (3) can be combined to produce:

$$\begin{aligned}
E_{new}[b_1|b_2] &= \frac{b_2 E[b_1|b_2] - E[b_1|b_2] E_{old}[b_1|b_2]}{b_2 - E[b_1|b_2]} \\
&= \frac{E[b_1|b_2] (b_2 - E_{old}[b_1|b_2])}{b_2 - E[b_1|b_2]}.
\end{aligned} \tag{2.4}$$

Equation (2.4) gives us the expected value of basal area at time 1 given basal area at time 2 for a tree in the “new” sample in terms of the same expected values for the population and the “old” sample. The expected value of b_1 given b_2 for the population is determined by a weighted regression over the trees in the “old” sample with the weights being the inverse probability of inclusion (i.e., multiply each observation by b_2/b_1).

We see that Equation (2.4) agrees with our intuitive expectation because, if:

$$\frac{b_2 - E_{old}[b_1|b_2]}{b_2 - E[b_1|b_2]} \leq 1$$

as we would expect since these trees were larger than the average tree at time 1 (selected proportional to b_1), then so too is

$$E[b_1|b_2] \geq E_{new}[b_1|b_2].$$