A FORM OF TWO-PHASE SAMPLING UTILIZING REGRESSION ANALYSIS

Michael A. Fiery and John R. Brooks[1]

Abstract—A two-phase sampling technique was introduced and tested on several horizontal point sampling inventories of hardwood tracts located in northern West Virginia and western Maryland. In this sampling procedure species and dbh are recorded for all "in-trees" on all sample points. Sawlog merchantable height was recorded on a subsample of intensively measured (second phase) sample points and these heights were predicted on the non-intensive (first phase) sample points. Regression analysis was used to predict heights on first phase points in order to achieve an estimate of board foot volume per acre for every point. Results indicate an improved estimate of the mean volume per acre when compared to traditional double sampling using basal area as the auxiliary variable. An unbiased sampling error was also achieved in this process.

INTRODUCTION

One of the major influences on forest inventory over the last few decades has been the desire to reduce field data collection time without sacrificing the accuracy and precision of sample based estimates of trees per acre, basal area, weight, and volume. The switch to the point sampling system, introduced by Grosenbaugh in late 1950s, was fueled by the obvious time savings as fewer "in-trees" were measured per sampling unit. Over the last 30 years there has been a slow migration to using larger basal area factors (BAFs) in sawtimber inventories, spurred by the empirical evidence that it provides less biased estimates of stand volume, but more likely due to the fact that fewer "in-trees" would be measured thus saving field data collection time (Wiant and others 1984, Brooks and McGill 2004). During this same period there was a parallel reduction in fixed area plot size from 0.25 and 0.20 acre plots to those of 0.1 acre in size. In the 1960s, a sampling technique commonly referred to double sampling was introduced by Freese (1962).

This sampling technique was developed to take advantage of the relationship between the variable of interest and some easily measured and highly correlated auxiliary variable so that only a subset of the overall sampling units would be intensively measured. One drawback to double sampling was that since dbh and species are only recorded on second-phase (intensive) points, no direct method of creating a stand and stock was available, though procedures were developed for their estimation (Matney and Parker 1991, Shiver and Borders 1996). Should an inventory require more accurate stand and stock tables, there are ways to do this without intensively measuring every tree on all sampling units. An inventory system can be designed where dbh, product, and species are tallied on all points, and tree heights are only measured on a subsample of these plots. Under the proposed sampling system, dbh, species, and sawlog merchantable heights would only be measured on second-phase (intensive) points. On all first-phase (non-intensive) samples, only dbh and species would be recorded. Using regression analysis, all heights necessary for volume estimation would be predicted on a species or species group basis. While this process requires more time than the use of traditional double sampling, it would be more efficient than measuring heights on all sampling units. This design would permit an accuracy equivalent to the intensive measurement of all sampling units for stems and basal area per acre. In areas where there are large variations in value based on species and size, diameter distribution data becomes increasingly important and may warrant the additional field inventory time. Although this technique has been employed in the South, no published record of the effects of height prediction on the accuracy and precision of typical volume sampling statistics has been found. The research that follows will:

[1] Michael A. Fiery, Graduate Student, and John R. Brooks, Associate Professor, West Virginia University, Forest Biometrics, Morgantown, WV 26506–6125.

1. Outline an inventory system where board foot volume is known (measured) on intensive points and estimated on all non-intensive points based on regression analysis to estimate sawlog merchantable height.

2. Through computer simulation, evaluate the behavior of the mean volume per acre and associated sampling error.

PROCEDURES

Datasets from several areas in West Virginia and Maryland were available for analysis in this study. Each dataset included measurements of species, dbh, and sawlog merchantable height on every point which permitted the comparison of both two-phase sampling methods (double sampling and height regression sampling) to estimates where all "in-trees" were intensively measured on all sampling units. The WVU Research Forest, Coopers Rock State Forest and the Coopers Rock Annex datasets are based on a 1999-2000 inventory conducted at both Coopers Rock State Forest and the West Virginia University Research Forest located in Monongalia and Preston Counties, WV. Primary species found in this inventory included yellow-poplar (*Liriodendron tulipifera* L.), northern red oak (*Quercus rubra* L.), red maple (*Acer rubrum* L.), chestnut oak (*Q. prinus* L.), and black cherry (*Prunus serotina* Ehrh.). These datasets were based on a BAF 20 point sampling inventory on a systematic grid. The Compartment 14, 1967 Single Species and Trout Pond datasets were also collected on the West Virginia University Research Forest as part of other research projects. The Tygart dataset was collected in the summer of 2004 from the Tygart Tract located in Dailey, WV. The tract is approximately 10 miles south of Elkins, WV and approximately 426 acres were inventoried. Primary species consisted of red maple, northern red oak, and chestnut oak. The original dataset consisted of 67, 1/5 acre circular plots where species, dbh (nearest 0.1 inch), sawlog merchantable height (0.1 foot) and total height (0.1 foot) were measured. Horizontal distance from plot center to every "in-tree" was also recorded to the nearest foot using an Impulse laser. The Savage River dataset comes from the Savage River State Forest in Garrett County, MD courtesy of the Maryland Forest Service. Primary species consisted of red oak, red maple, and chestnut oak. This dataset consisted of 214, 1/5 acre circular plots which formed the basis of a continuous forest inventory system located throughout the 53,473 acre forest. At each plot, species, dbh (nearest 0.1 inch), and number of 8-foot logs were tallied for every "in-tree". Horizontal distance from plot center to every "in-tree" was also recorded to the nearest foot using an Impulse laser. All datasets originally based on fixed area plots included accurate measurements of horizontal distance to each "in-tree". These datasets were converted to point sample inventories based on a BAF of 20. All trees having a horizontal distance from point center equal to or less than the critical distance for that tree size were included in the final dataset.

Since all datasets were originally based on the intensive measurement of all "in-trees" on all points, board foot volume of every "in-tree" was calculated based on field measurements of dbh and sawlog merchantable height and the board foot volume equations published by Scott (1979) for International ¼ inch log rule. In this study, the "actual" mean volume per acre used for comparison purposes is based on simple random sampling statistics of this horizontal point sample data.

To conduct the double sample point sample inventory, the existing datasets were sampled using a 1:4 ratio where one out of every four samples points was selected as an intensively measure point utilizing the recorded species, dbh and sawlog merchantable height data. On all other points, only species and dbh were utilized. A ratio of means estimator was used to calculate mean board foot volume per acre and the associated standard error employing basal area as the auxiliary variable (Shiver and Borders 1996).

Height regression sampling was applied to the same samples selected in double sampling inventory to minimize variation between the two systems. In this case, the same points that were selected as second phase sampling units (intensively measured sample points) in the double sample inventory were also used as a basis for the merchantable height regressions. Under this two-phase sampling system, species, dbh and sawlog merchantable height were utilized on intensive points while only species and dbh were used

on non-intensive points. Regression analysis was used to predict sawlog merchantable heights on all non-intensive points, thus providing the necessary information to calculate boardfoot volume for every sample point. A common height model mentioned by Avery and Burkhart (2002) was used and is of the form:

$$Ln(MHT) = \beta_0 + \beta_1\left(\frac{1}{DBH}\right) + e_i \tag{1}$$

where

DBH = diameter at breast height (inch)
MHT = sawlog merchantable height (feet)
β_0, β_1 = parameters to be estimated from the data
e_i = error (feet)

Separate parameters were established for each species whenever a sample size of five or more was available. If less than five observations were available, the species was grouped in the "all other" category.

Two different methods for data analysis were conducted and evaluated for height regression sampling, each producing a different set of results. Sampling error for Method 1 is easily calculated but not statistically sound while Method 2 provides a more rigorous approximation of the sampling error.

Method 1 (SRS method) for height regression sampling calculates inventory statistics for board foot volume per acre using simple random sampling techniques. The assumption is made that the error associated with predicted heights on non-intensive (first phase) points is minimal and that the estimates of volume per acre and the associated standard error, can be estimated using simple random sampling statistics.

Method 2 (ratio method) for height regression sampling uses an estimate of volume (using predicted heights for all trees) for each point. The actual volume on just the intensively measured points is based on measured tree heights. At this point the dataset can be treated as a double sample where actual volume is the variable of interest and estimated volume is the auxiliary variable. The mean volume per acre and standard error can be calculated using equations (2) and (4) respectively. A ratio estimator was used to find the mean volume per acre:

$$\bar{y}_{hrs} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \times \frac{\sum_{i=1}^{n'} x_i}{n'} = \hat{R}\bar{x}' \tag{2}$$

where

$\bar{y}_{hrs}$ = mean volume per acre
y_i = actual volume per acre on intensive points
x_i = predicted volume per acre on intensive/non-intensive points
n = number of intensive points
n' = total number of points

The variance of the ratio can be calculated from the equation:

$$S_{\hat{R}}^2 = \frac{\sum_{i=1}^{n} y_i^2 - 2\hat{R}\sum_{i=1}^{n} x_i y_i + \hat{R}^2 \sum_{i=1}^{n} x_i^2}{n-1} \tag{3}$$

where

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

The overall variance can be calculated by (Shiver and Borders 1996):

$$S^2_{\bar{y}_{hrs}} = \frac{S^2_y}{n'} + \frac{S^2_{\hat{R}}}{n}\left(\frac{n' - n}{n'}\right) \quad (4)$$

In order to investigate the estimation properties of both the mean and standard error for the sampling techniques described, a Visual Basic 6.0 simulation program was written to resample existing datasets where second phase sample points were selected at random (without replacement). This procedure was employed for both double sampling and height regression sampling (methods 1 and 2). For each simulation, a 1:4 ratio of intensive to non-intensive samples was employed where all intensive (second phase) points were selected using a random number generator, thus providing unique inventories each simulation. A total of 500 simulations were conducted for each dataset, providing estimates of the mean volume per acre and the standard error for each of the 500 simulations.

RESULTS

For each of the 8 inventory datasets, mean board foot volume per acre was estimated using measured diameters and sawlog merchantable heights that were available for every "in-tree" in the dataset. The simple random sampling statistics based on these measurements are considered the actual volumes for this study. One out of every four points was then selected to be used as an intensively measured sample point (second phase) for both a traditional double sample and the proposed height regression sampling technique. For the intensively measured points, species, dbh and sawlog merchantable height measurements were utilized. For the non-intensive (first phase) sample points, only species and dbh information was used. Table 1 includes the mean board foot volume estimate based on the simple random sample mean (SRS) from the assumed "actual" volume, the mean based on a double sample ratio of means estimator (DS) and two estimates based on the use of height regression sampling (HRS) using the two estimation methods described previously. In five of the eight inventories, the mean board foot volume based on the HRS method 1 procedure was more accurate than the common double sampling approach using basal area as the auxiliary variable (table 1). While in all but one of the inventories, the mean board foot volume based on the HRS method 2 procedure was more accurate than the double sampling approach (table 1). The results of the regression process indicate that some increase in variance with increasing tree size occurred, but the distribution of merchantable height errors appeared random (fig. 1). In most cases, predicted heights were within 10 feet of the actual height of each tree at least 60 percent of the time (table 2). The estimated standard errors for each sampling scheme are shown in table 3. In each of the eight inventories, both HRS method procedures resulted in an estimate of the standard error that were closer to that based on the complete measurement of every sample tree (SRS). Both the DS and HRS systems used the same 1:4 ratio of second phase to first phase samples and all intensively measured points (second phase) were the same sample points in both instances.

The results from the 500 simulations conducted on each of the 8 inventory datasets indicate that the HRS method 2 estimates of the mean board foot volume per acre had a lower RMSE in all cases and a smaller average bias in 5 of the 8 inventories tested (table 4). Only one of the HRS method 1 estimates had a smaller average bias but all eight inventories still had a lower RMSE. Both the DS and HRS simulations provided what appeared to be unbiased estimates of the mean board foot volume per acre, with both HRS procedures showing a higher level of precision (fig. 2). The estimates of the sampling error for the

Table 1—The mean and bias in board foot volume per acre for inventory tract and sampling scheme

			Mean board foot volume per acre				Difference from SRS		
Tract	Acres	Points	SRS	DS	HRS (srs)	HRS (ratio)	DS	HRS (srs)	HRS (ratio)
Tygart	426	67	7,883.4	7,935.5	7,834.8	7,840.7	52.1	-48.6	-42.7
WVU Research Forest	7,594	2,013	9,676.1	9,730.9	9,520.8	9,640.9	54.7	155.4	-35.3
Coopers Rock State Forest	4,037	1,081	10,643.7	10,426.8	10,589.0	10,743.4	-217.0	-54.8	99.7
Coopers Rock Annex	364	98	10,341.9	9,945.2	10,511.0	10,418.4	-396.8	169.0	76.5
Compartment 14	138	52	11,076.6	10,909.2	11,209.8	11,053.8	-167.4	133.2	-22.7
1967 single species	3,500	384	2,539.4	2,531.3	2,473.0	2,546.5	-8.1	-66.3	7.1
Trout Pond	N/A	30	14,591.6	14,988.3	13,899.0	13,718.9	396.5	-692.9	-872.9
Savage River	53,473	214	10,363.2	8,359.0	10,551.4	10,732.6	-2,004.2	188.2	369.4

SRS = simple random sample mean; DS = double sample ratio of means estimator; HRS = height regression sampling.

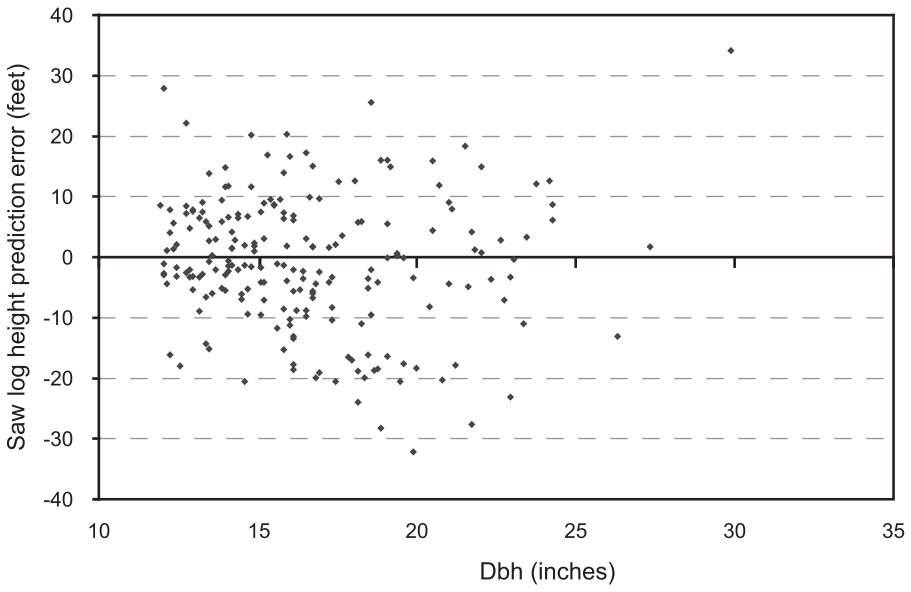

Figure 1—Height prediction error (foot) by d.b.h. across all species (Tygart Tract).

Table 2—Percentage of predicted saw log heights within 10 percent of the actual measured height by inventory tract

Tract	n	Within 10 feet	Percent within 10 feet
Tygart	210	144	68.57
WVU Research Forest	8,052	5,331	66.21
Coopers Rock State Forest	4,892	2,158	44.11
Coopers Rock Annex	392	240	61.22
Compartment 14	254	158	62.20
1967 single species	1,253	966	77.09
1967 species specific	331	261	78.85
Trout Pond	178	132	74.16
Savage River	932	630	67.60

Table 3—Standard error by inventory tract and sampling scheme

	Standard error (board foot per acre)				Standard error of the mean			
Tract	SRS	DS	HRS (srs)	HRS (ratio)	SRS	DS	HRS (srs)	HRS (ratio)
					percent			
Tygart	683.8	1,055.2	659.2	662.0	8.7	13.3	8.5	8.4
WVU Research Forest	153.9	230.6	145.5	159.7	1.6	2.4	1.7	1.7
Coopers Rock State Forest	200.1	343.4	192.4	214.6	1.9	3.3	2.0	2.0
Coopers Rock Annex	653.2	911.9	626.2	670.5	6.3	9.2	6.4	6.4
Compartment 14	916.3	1,534.8	894.6	890.7	8.3	14.1	8.0	8.1
1967 single species	130.9	152.3	119.9	146.7	5.2	6.0	5.9	5.8
Trout Pond	1,254.8	1,717.8	1,190.8	1,354.4	8.6	11.5	9.7	9.9
Savage River	481.0	1,840.8	480.4	502.2	4.6	22.0	4.8	4.7

SRS = simple random sample mean; DS = double sample ratio of means estimator; HRS = height regression sampling.

Table 4—Average bias and root mean squared error for volume estimates by inventory tract and sampling scheme (based on 500 simulations)

			Average bias (board foot per acre)			RMSE (board foot per acre)		
Tract	Acres	Points	DS	HRS (srs)	HRS (ratio)	DS	HRS (srs)	HRS (ratio)
Tygart	426.0	67.0	82.4	69.5	33.5	952.9	239.8	223.7
WVU Research Forest	7,594.0	2,013.0	8.3	-125.0	3.0	197.1	137.8	57.0
Coopers Rock State Forest	4,037.0	1,081.0	-15.6	-146.0	10.8	266.0	168.8	82.0
Coopers Rock Annex	364.0	98.0	39.0	163.7	74.5	787.4	326.3	286.0
Compartment 14	138.0	52.0	75.3	189.5	198.3	1,413.3	494.1	471.0
1967 single species	3,500.0	384.0	-3.4	-72.7	-3.3	75.0	90.0	52.5
Trout Pond	N/A	30.0	-330.7	324.0	158.6	1,290.8	613.1	524.2
Savage River	53,473.0	214.0	20.8	-192.9	-55.9	1,283.3	301.4	237.8

DS = double sample ratio of means estimator; HRS = height regression sampling; SRS = simple random sample; RMSE = root mean squared error.

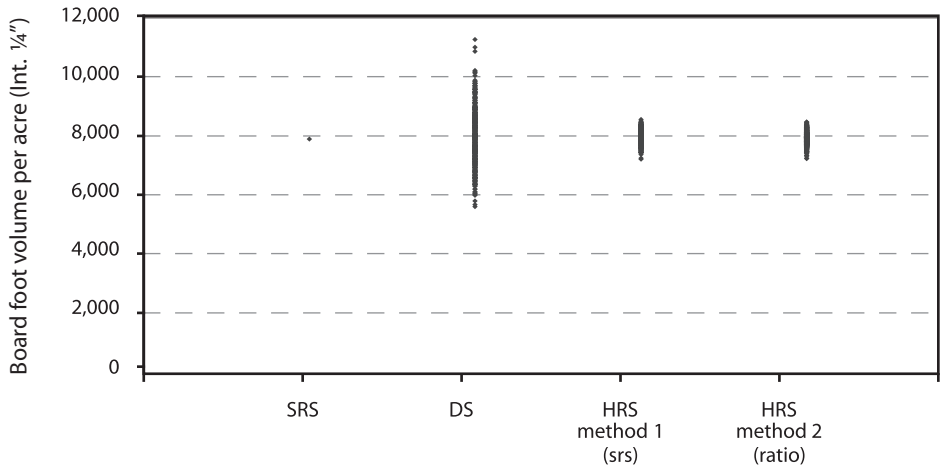

Figure 2—Variation in mean board foot volume per acre by sampling type for the Tygart Tract (based on 500 simulations).

HRS procedures were centered around the SRS estimate an appeared less variable than the traditional DS estimates (fig. 3). The DS estimates appeared to be biased in a positive direction. The relationship between actual and predicted volume, which was used to estimate the HRS sampling error, had a correlation coefficient of 0.999 and this relationship is depicted in figure 4.

DISCUSSION

The overall effect on mean volume per acre when some sampling units have measured heights and others have estimated heights is unknown. The variance is most likely reduced as the natural variation in heights by diameter and thus on volume has been removed through the prediction process.

The use of regression analysis to predict merchantable heights to obtain volume estimates on a subset of sampling units has provided some positive results. Mean volume estimation was usually more accurate

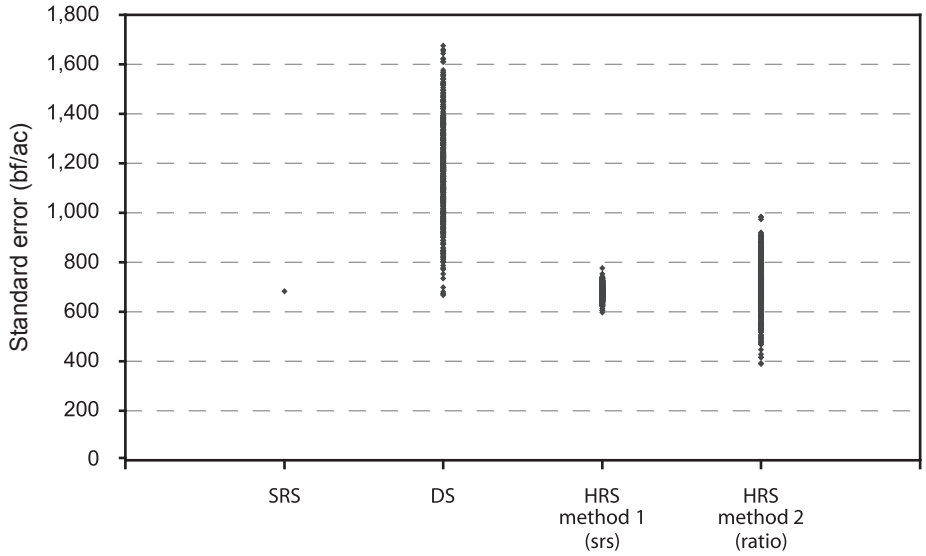

Figure 3—Variation in the standard error (board feet per acre) by sampling type for the Tygart Tract (based on 500 simulations).

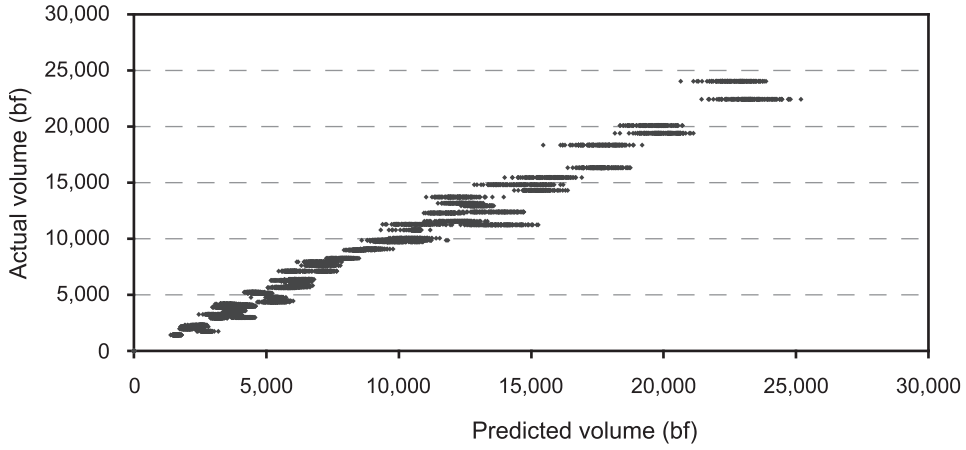

Figure 4—Relationship between actual and predicted volumes based on estimated heights for phase 2 points on the Tygart Tract (based on 500 simulations).

and precise than the traditional DS procedures. This study employed two different methods for achieving an estimate of sampling error. Method 1 employed a simple random sampling estimate of variance ignoring the fact that sawlog merchantable heights were predicted on all non-intensive (first phase) samples. In HRS method 2, the analysis was reformulated as a true double sampling application where a ratio estimator was used to estimate the sampling error. For most of the inventories investigated, both HRS sampling methods provided a smaller standard error than traditional DS approach. It appears that both HRS sampling methods provide positive results and although method 1 ignores the effect of height prediction on the overall variance, it still resulted in a seeming unbiased estimate of the mean board foot volume per acre and an unbiased, but slightly less variable, estimate of the sampling error.

The application of the HRS procedure is not warranted in many cases. The measurement of dbh on every sample tree may require more time resources than a single inventory can justify. However, in those situations where the additional diameter distribution data is desired, this procedure requires less field collection time than the complete enumeration of every sample point and provides reasonably accurate estimates of mean volume per acre even in variable hardwood populations.

LITERATURE CITED

Avery, T.E.; Burkhart, H.E. 2002. Forest measurements. McGraw-Hill Companies Inc. New York. pp.185-186.

Brooks, J.R.; McGill, D.M. 2004. Evaluation of multiple fixed area plot sizes and BAFs in even-aged hardwood stands. Yaussy, D.A., D.M. Hix, R.P. Long and G.P. Charles eds. Proceedings 14th Central Hardwood Forest Conference; 2004 March 16-19; Wooster, OH. Gen. Tech. Rep.NE-316. 94-100 p.

Freese, F. 1962. Elementary forest sampling. U.S. Department of Agriculture, Forest Service Agriculture Handbook 232.

Matney, T.G.; Parker, R.C. 1991. Stand and stock tables from double-point samples. Forest Science 37(6):1605-1613.

Scott, C.T. 1979. Northeastern forest survey board-foot volume equations. U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station, Research Note NE-271.

Shiver, B.D.; Borders, B.E. 1996. Sampling techniques for forest resource inventory. John Wiley & Sons Inc. New York.

Wiant, H.V., Jr.; Yandle, D.O.; Andreas, D. 1984. Is BAF 10 a good choice for point sampling? Northern Journal of Applied Forestry 1:23-24.