United States
Department of
Agriculture

Forest Service

Southeastern Forest
Experiment Station

Research Paper
SE-264

A Comparison of Regional and Site-Specific Volume Estimation Equations

Joe P. McClure
Jana Anderson
Hans T. Schreuder

A Comparison of Regional and Site-Specific Volume Estimation Equations

Joe P. McClure, Project Leader
Forest Inventory and Analysis
Southeastern Forest Experiment Station
Asheville, North Carolina

Jana Anderson, Statistician
Business School, Colorado State University
Fort Collins, Colorado

Hans T. Schreuder, Project Leader
Multiresource Inventory Techniques
Rocky Mountain Forest and Range Experiment Station
Fort Collins, Colorado

ABSTRACT

Regression equations for volume by region and site class were examined for loblolly pine. The regressions for the Coastal Plain and Piedmont regions had significantly different slopes. The results showed important practical differences in percentage of confidence intervals containing the true total volume and in percentage of estimates within a specific proportion of the true total in a simulation study. Sampling from a modified Coastal Plain population with the same diameter distribution as the Piedmont population and a modified Piedmont population with the same diameter distribution as the Coastal Plain population showed that having the proper diameter distribution did not improve predictions either in terms of confidence intervals or in getting more estimates within a specified percentage of true volume.

The regressions for site class populations had significantly different slopes. The simulation study showed that it mattered considerably which sites the samples were drawn from.

KEYWORDS: Mensuration, loblolly pine, practical differences, statistical differences, weighted regression equations.

There is a belief in forestry that trees of a given species require separate volume tables for different sites and regions of the country. This, despite the fact that volume regression equations of the type $V = a + b\ D^2H$, where V = volume and D^2H = diameter at breast height squared times height, tend to give very similar-looking estimated regression coefficients a and b for most data sets of a given species and across many species.

This Paper examines the gain associated with population-specific equations for merchantable cubic-foot volume inside bark of loblolly pine (*Pinus taeda* L.) by regions and sites. Data for the study are from the Southeastern Forest Experiment Station's Forest Inventory and Analysis (FIA) unit. Because the sample sizes are very large, we expected the difference in equations to be statistically significant. We therefore also tried to assess the practical importance of the differences.

Weighted simple linear regressions with assumed known weights k = 1.5 were used in all cases. Weighted regression is similar to simple linear regression but assumes an increase in variability in the variable of interest (volume) (measured by k) with an increase in the independent variable (D^2H). McClure and others[1] estimated k to be 1.5 for large white oak (*Quercus alba* L.) and loblolly pine data sets for the model:

$$V = \alpha + \beta\ D^2H + e \qquad (1)$$

where the mean value of e is 0 and the variance of e is $V_a(e) = \sigma^2(D^2H)^k$, where σ^2 is the mean square residual.

The literature is unclear on the need for separate equations for different regions or even different species. Gevorkiantz and Olsen[2] developed composite board-foot, cubic-foot, and cordwood volume tables combining conifer and hardwood species. The model used for cubic-foot volume was the special case of equation (1):

$$V = \frac{(0.42)\pi}{144(4)}\ D^2H$$

---

[1]McClure, Joe P.; Schreuder, Hans T.; Wilson, Rodney L. 1983. A comparison of several volume table equations for loblolly pine and white oak. Res. Pap. SE-240. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station. 8 pp.

[2]Gevorkiantz, S.R.; Olsen, L.P. 1955. Composite volume tables for timber and their application in the Lake States. Tech. Bull. 1104. Washington, DC: U.S. Department of Agriculture. 51 pp.

with D in inches and H in feet. The authors concluded that these tables were adequate for large tracts generally and for small tracts for reconnaissance surveys or for timber of low value. They also pointed out that the definition of a good model is highly subjective.

Van Deusen and others[3] compared volume equations for old-field loblolly pine plantations for 702 trees from the Georgia Piedmont; 300 trees from Alabama, Arkansas, and Mississippi Gulf Coastal Plains; 378 trees from the Piedmont and Coastal Plain of Virginia, Delaware, Maryland, and North Carolina; and 340 trees from the Tennessee, Alabama, and Georgia highlands. They used the combined variable equation

$$V = a + b\ D^2H$$

based on model (1) with k = 2. They eliminated the influence of diameter distribution on the equation comparison by using the Mississippi diameter distribution as a model and modifying the other data sets to yield the same diameter distribution. They found statistically significant differences (at the 0.05 level) in outside-bark volume equations and attribute them to differences in measurement technique. The question of practical significance remained.

Methods

In our analyses, we addressed the subject of practical significance. If a population-specific equation does not have to be used, inventory costs can be reduced because tree volumes generally do not have to be measured separately by region. Our criteria for determining the practical importance of differences in models were:

1. The relative proportions of 95-percent confidence intervals containing the total volume of a population (region or site class) for samples from that population (say Coastal Plain) or a different population (say Piedmont).

2. Relative proportions of regression estimates that are within a set percentage (1, 2, 3, 4, and 5 percent) of the total volume for a population (region or site class) for samples from that population or a different but similar population.

Data Sets

The loblolly-tree-volume data base described by McClure and others[1] was used as a starting point in our study. Total volume in cubic feet, total height in feet, and diameter in inches were measured. These trees were selected to ensure a large sample of trees over the range of volumes and D^2H values. Numbers of trees by size and site class, therefore, are not representative of the populations of interest, which are the actual loblolly pine populations in the Southeast. Since there were only 75 trees in the mountain region and 55 in site class 5, the comparisons were limited to the Piedmont (P) and Coastal Plain (C) regions and productive sites (S_{1-3}) and unproductive sites (S_{4-5}).[4] Estimates of 2-inch diameter distributions for these two regions and two site classes are available from the very large permanent-plot data base maintained by the Southeastern Forest Experiment Station FIA unit. One-inch diameter class frequencies were obtained by interpolation. The tree-volume data base was then modified by random addition and subtraction of trees in the proper diameter classes to yield data sets with about the same diameter frequencies as the interpolated estimates.

[3]Van Deusen, Paul C.; Sullivan, Alfred D.; Matney, Thomas G. 1981. A prediction system for cubic foot volume of loblolly pine applicable through much of its range. Southern Journal of Applied Forestry 5(4):186-189.

[4]Production classes are based on cubic feet of yield for fully stocked natural stands at culmination of mean annual increment. Sites 1 to 3 are those growing more than 85 cubic feet, site 4 those with 50 to 85 cubic feet, and site 5 those with 20 to 50 cubic feet per acre per year.

The resulting four large samples (two regions and two site classes) were treated as populations for purposes of this study, and two new data sets, P* and C*, were generated. P* consisted of Piedmont trees with the same diameter distribution as the Coastal Plain, and C* of Coastal Plain trees with the same diameter distribution as the Piedmont population.

Results

Regional and site class comparisons will be discussed separately. For each region or site class, the prediction equation

$$\hat{V} = N\ (a + b\ \overline{D^2H})$$

was used, where N is total number of trees and $\overline{D^2H}$ is the mean for the population (region or site) and a and b are regression coefficients estimated from the sample drawn.

Regional Comparisons

Coastal Plain (C) and Piedmont (P) populations consisted of 1,801 and 1,800 trees, respectively. Skewness (b_1) and kurtosis ($\sqrt{b_2}$) for volume are b_1 = 3.1 and $\sqrt{b_2}$ = 14.6 for C, b_1 = 5.5 and $\sqrt{b_2}$ = 61.9 for P, b_1 = 3.7 and $\sqrt{b_2}$ = 22.6 for C*, and b_1 = 5.2 and $\sqrt{b_2}$ = 46.4 for P*. These values show that the populations are highly skewed; there are many trees with small and few with large volumes.

There were two analyses: one for statistical significance and one for practical utility. The first question was whether the weighted regression equations with known weights k = 1.5 for the two regions,

$$V_C = -0.734 + 0.00213\ D^2H \qquad (2)$$

with 1,801 trees and mean square error = 0.00001 and

$$V_P = -0.769 + 0.00208\ D^2H \qquad (3)$$

with 1,800 trees and mean square error = 0.00001 are significantly different or can be combined into one general equation

$$V_R = -0.764 + 0.00211\ D^2H \qquad (4)$$

with 3,601 trees and mean square error = 0.00001.

Because of the large sample sizes involved, the regression equations are significantly different. The 95-percent confidence interval constructed around the difference between the slopes of the two regional regressions is the interval (0.00002316, 0.00008105) which does not contain 0. The regression equations appear to be very similar, and the question of practical importance of differences remained.

The weighted Coastal Plain regression equation based on C* is

$$V_{C*} = -0.688 + 0.00211\ D^2H \qquad (5)$$

with 1,800 trees and mean square error = 0.00001. The Piedmont equation based on P* is

$$V_{P*} = -0.734 + 0.00206\ D^2H \qquad (6)$$

with 1,801 trees and mean square error = 0.00001.

The second question is: how well do the regional regression equations predict total volume for each region?

Because the test is limited to regression models, average D^2H per tree and number of trees N in the region for which predictions were being made were assumed known and used in both equations. One thousand samples of sizes 20, 40, 60, and 80 were drawn at random from each region. Total volume of the region was predicted for each sample by computing the regression coefficients for each sample and combining this with known N and $\overline{D^2H}$ for the region. The proportion of 95-percent sample confidence limits containing the actual total volume for the region was computed along with the percentage of volume estimates within 1, 2, 3, 4, and 5 percent of the actual total volume.

The confidence interval results are shown in table 1. Samplings from either C or C* are better in predicting volume for C then sampling from P or P*. When P or P* is used rather than C, there is a loss from 15.4 to 16.9 percent for samples of size 20, and from 53.6 to 56.0 percent for samples of size 80. Whether P or P* is used seems to make little difference. Using C* rather than C actually improves confidence interval estimation when predicting volume for C.

When predicting total volume for P, using C or C* results in slightly poorer confidence intervals than using P or P*. The loss ranges from 2.3 to 2.4 percent for samples of size 20 and from 5.7 to 8.5 percent for samples of size 80. There seems to be a slight edge in using P* rather than P when predicting P volume.

The percentages of time that the estimated volume is within 1, 2, 3, 4, and 5 percent of true volume for the predicted region are shown in table 2 for n = 20 and n = 40 and in table 3 for n = 60 and n = 80. Using samples for C* rather than C to predict volume for C gives inconsistent results. For n = 20 and n = 60, sampling C* is somewhat less successful and for n = 40 and n = 80 somewhat more successful than sampling C. Sampling from P* in predicting volume for C is quite unsuccessful relative to sampling C or C*. Sampling from P* is only slightly better than sampling from P. Having a population with the same diameter distribution as the population being sampled results in only slightly improved estimates.

When predicting volume of P, sampling from P* rather than from P results in higher percentages of times that the estimated volume is within 1, 2, 3, 4, or 5 percent of the true volume even though P* has a different diameter distribution than P. Sampling from C or C* is clearly worse than sampling from either P or P*. There is little difference in the results from C or C*, even through C* has the same diameter distribution as P.

Site Comparisons

For site class comparisons, there were 1,798 trees in sites 1-3 (S_{1-3}) and 1,401 trees in sites 4-5 (S_{4-5}). Skewness (B_1) and kurtosis ($\sqrt{b_2}$) for volume are b_1 = 4.79 and $\sqrt{b_2}$ = 40.47 for S_{1-3} and b_1 = 2.91 and $\sqrt{b_2}$ = 14.83 for S_{4-5}.

Table 1.--Proportion of confidence intervals containing actual total volume for a region for samples of sizes 20, 40, 60, and 80 from each region and percentage loss in this proportion resulting from sampling in the other region. Each value is based on 1,000 iterations.

Population sampled	Region C				Region P			
	n = 20	n = 40	n = 60	n = 80	n = 20	n = 40	n = 60	n = 80
				PROPORTION				
C	0.902	0.897	0.900	0.893	0.878	0.827	0.801	0.733
P	0.750	0.612	0.495	0.414	0.899	0.866	0.823	0.801
C*	0.926[a]	0.928	0.902	0.915	0.877	0.841	0.817	0.755
P*	0.763	0.624	0.493	0.393	0.896	0.871	0.878	0.872
				PERCENTAGE LOSS				
Other region	16.9	31.8	45.0	53.6	2.3	4.5	2.7	8.5
C*	0.0	0.0	0.0	0.0	2.4	2.9	0.7	5.7
P*	15.4	30.4	45.2	56.0	0.3	0.0	0.0	0.0

[a]If sampling from the other region yields a better result, the percentage loss is set at 0.0 percent.

Table 2.--Percentage of regression estimates that are within 1, 2, 3, 4, and 5 percent of true total volume for a region and the relative success rates for samples of sizes 20 and 40 drawn from regions C, C*, P, and P*

Region predicted for--	Percent deviation allowed	Region sampled (n=20)				Success rate			Region sampled (n=40)				Success rate		
		C	C*	P	P*	C*	R[a]	P*	C	C*	P	P*	C*	R[a]	P*
		Percent													
Coastal Plain	1	19.8	18.5	11.2	9.5	93.4	56.6	48.0	23.5	27.5	10.8	10.4	100.0	46.0	44.3
	2	37.3	35.5	23.5	19.7	95.2	63.0	52.8	46.3	52.3	20.3	23.3	100.0	43.8	50.3
	3	53.8	52.5	34.0	33.9	97.6	63.2	63.0	66.4	72.5	32.3	38.0	100.0	48.6	57.2
	4	67.8	66.4	45.0	48.2	97.9	66.4	71.1	81.1	83.5	45.4	52.8	100.0	56.0	65.1
	5	77.2	77.8	56.6	60.7	100.0	73.3	78.6	90.1	91.7	58.5	67.0	100.0	64.9	74.4
Piedmont	1	17.4	16.0	20.2	21.1	86.1	79.2	100.0	22.2	21.3	24.0	26.2	88.7	92.5	100.0
	2	35.6	32.7	38.8	40.1	91.8	84.3	100.0	41.6	40.7	45.5	49.3	89.5	91.4	100.0
	3	49.5	47.3	53.2	54.8	93.0	88.9	100.0	59.9	58.8	64.1	68.0	91.7	93.4	100.0
	4	63.1	60.1	66.7	69.4	94.6	90.1	100.0	74.3	74.0	76.6	82.4	96.6	97.0	100.0
	5	73.1	71.7	75.7	79.6	96.6	94.7	100.0	84.0	83.5	86.3	91.7	96.8	97.3	100.0

[a]Ratio of successful predictions in the unsampled region to successful predictions in the sampled region in percent.

Table 3.--Percentage of regression estimates that are within 1, 2, 3, 4, and 5 percent of true total volume for a region and the relative success rates for samples of sizes 60 and 80 drawn from regions C, C*, P, and P*

Region predicted for--	Percent deviation allowed	Region sampled (n=60)				Success rate			Region sampled (n=80)				Success rate		
		C	C*	P	P*	C*	R[a]	P*	C	C*	P	P*	C*	R[a]	P*
		Percent													
Coastal Plain	1	35.1	31.5	6.6	8.6	89.7	18.8	24.5	37.3	36.6	6.0	7.2	98.1	16.1	19.3
	2	62.2	59.1	17.6	20.4	95.0	28.3	32.8	66.4	64.4	16.3	17.7	97.0	24.5	26.7
	3	80.3	77.5	30.3	35.0	96.5	37.7	43.6	84.4	85.9	30.9	34.2	100.0	36.6	40.5
	4	90.7	89.5	45.7	54.5	98.7	50.4	60.1	93.6	94.2	48.9	55.2	100.0	52.2	59.0
	5	96.5	95.5	64.2	72.0	99.0	66.5	74.6	97.6	98.1	67.1	71.7	100.0	68.7	73.5
Piedmont	1	23.3	23.3	27.0	31.0	86.3	86.3	100.0	21.6	22.0	27.0	32.7	81.5	80.0	100.0
	2	46.8	45.1	50.4	57.5	92.9	89.5	100.0	43.3	44.5	50.6	65.6	87.9	85.6	100.0
	3	65.2	66.9	68.9	78.6	94.6	97.1	100.0	64.6	66.4	69.5	83.7	95.5	92.9	100.0
	4	80.1	81.0	82.0	91.1	97.7	98.8	100.0	81.9	82.4	83.7	95.0	98.4	97.8	100.0
	5	88.5	90.7	89.6	97.0	98.8	100.0	100.0	92.0	91.4	92.8	98.9	98.5	99.1	100.0

[a]Ratio of successful predictions in the unsampled region to successful predictions in the sampled region in percent.

Equations S_{1-3} were compared with equations S_{4-5} in the same way as those for regions were compared. The better sites, S_{1-3}, were combined because sample sizes for the individual sites were inadequate. The site equations are:

$$V_{S_{1-3}} = -0.791 + 0.00214\ D^2H \qquad (7)$$

with 1,798 trees and mean square error = 0.00001

$$V_{S_{4-5}} = -0.191 + 0.00217\ D^2H \qquad (8)$$

with 1,401 trees and mean square error = 0.00001

A general weighted regression for all trees is

$$V_S = -0.539 + 0.002160\ D^2H \qquad (9)$$

with 3,199 trees and mean square error = 0.00001.

The regression equations of S_{1-3} and S_{4-5} are significantly different because the 95-percent confidence intervals constructed around the difference of the slopes of the two regional regressions is the interval (-0.00005917, -0.00000172) which does not include 0.

Because the test is limited to regression models, average D^2H per tree and number of trees N in the site class for which predictions were being made were assumed known and used in all prediction equations. One thousand samples of sizes 20, 40, 60, and 80 were drawn at random from each site class. Total volume of each site class was predicted for each sample by computing the regression coefficients and combining this with N and $\overline{D^2H}$ for each site class. The percentages of 95-percent sample confidence limits containing the actual total volume for each site class were computed along with the percentages of volume estimates within 1, 2, 3, 4, and 5 percent of the actual total volume.

Table 4 shows that there is a loss in sampling the wrong site class in terms of percentage of confidence intervals containing the true total volume. The loss increases with increase in sample size, going from 24.6 to 83.6 percent for n = 20 to n = 80 when sampling for S_{1-3} and from 17.8 to 65.6 percent for n = 20 to n = 80 when sampling for S_{4-5}.

The percentages of times that the estimated volume is within 1, 2, 3, 4, and 5 percent of true volume for the predicted site are shown in table 5 for n = 20, n = 40, n = 60, and n = 80. It is clearly critical to use the regression from the correct site class to predict total volume. Predicting for S_{1-3} with the regressions from S_{4-5} results in success rates of 53.4 percent or less for n = 20 to 37.8 percent or less for n = 80 with the 1, 2, 3, 4, and 5-percent deviations allowed relative to the success rate for S_{1-3} regressions. Similarly, predicting for S_{4-5} with regressions from S_{1-3} results has success rates of 75.4 percent or less for n = 20 to 65.9 percent or less for n = 80 relative to the success rate for S_{4-5} regressions.

Conclusions

Although the results are subjective because the practical utility criterion used differs between practitioners, firm conclusions can be drawn. A considerable price is paid by sampling from the other regions even when population size and average $\overline{D^2H}$ are known for a population. Although regression equations for regions seemed to differ very little, they had significantly different slopes that yielded differences of practical importance. Percentage of confidence intervals containing the parameter of interest and percentage of estimates close to this parameter were considerably better when samples were drawn from the proper region. Having the proper diameter distribution did not help when drawing samples from the other region.

For site class populations, the regressions were significantly different and there was considerable difference in terms of percentage of confidence intervals containing the parameter of interest and percentage of estimates close to this parameter.

Table 4.--Percentage of confidence intervals containing actual total volume for a site for samples of sizes 20, 40, 60, and 80 from each site and percentage loss in this proportion resulting from sampling the other site class. Each value is based on 1,000 iterations.

Site class sampled from--	Predictions							
	Sites 1-3				Sites 4-5			
	n = 20	n = 40	n = 60	n = 80	n = 20	n = 40	n = 60	n = 80
	PROPORTION							
Sites 1-3 equation	0.907	0.892	0.902	0.885	0.735	0.564	0.421	0.306
Sites 4-5 equation	0.684	0.452	0.268	0.145	0.894	0.921	0.907	0.890
	PERCENTAGE LOSS							
Loss from sampling other site class	24.6	49.3	70.3	83.6	17.8	38.8	53.6	65.6

Table 5.--Percentage of regression estimates that are within 1, 2, 3, 4, and 5 percent of true total volume for a region for samples of sizes 20, 40, 60, and 80 drawn from two site classes

Sites predicted for--	Percent deviation allowed	n = 20 Site sampled from--			n = 40 Site sampled from--			n = 60 Site sampled from--			n = 80 Site sampled from--		
		S_{1-3}	S_{4-5}	Success rate[a]	S_{1-3}	S_{4-5}	Success rate[a]	S_{1-3}	S_{4-5}	Success rate[a]	S_{1-3}	S_{4-5}	Success rate[a]
S_{1-3}	1	19.4	5.4	27.8	27.6	4.4	15.9	34.7	2.9	8.4	40.8	0.9	2.2
	2	39.8	14.5	36.4	52.7	10.5	19.9	60.3	5.8	9.6	70.1	3.3	4.7
	3	57.0	24.3	42.6	72.7	18.9	26.0	80.8	12.8	15.8	85.3	8.1	9.5
	4	70.1	33.5	47.8	85.1	30.1	35.4	92.8	23.5	25.3	94.6	19.4	20.5
	5	80.1	42.8	53.4	93.1	43.0	46.2	96.9	39.6	40.9	98.6	37.3	37.8
S_{4-5}	1	10.3	18.9	54.5	8.9	27.9	31.9	5.9	30.4	19.4	4.3	34.8	12.4
	2	22.5	37.4	60.2	19.8	50.7	39.1	15.2	59.2	25.7	13.0	64.2	20.2
	3	34.3	53.1	64.6	32.4	70.9	45.7	29.6	78.5	37.7	25.3	82.1	30.8
	4	47.1	66.5	70.8	46.6	83.5	55.8	46.1	90.9	50.7	44.6	93.3	47.8
	5	57.7	76.5	75.4	61.4	92.6	66.3	64.6	96.9	66.7	64.4	97.7	65.9

[a]Ratio of successful predictions of the other site class to successful predictions of the correct site class population in percent.

McClure, Joe P.; Anderson, Jana; Schreuder, Hans T.

A comparison of regional and site-specific volume estimation equations. Res. Pap. SE-264. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1987. 9 pp.

Regression equations for volume by region and site class were examined for loblolly pine. The regressions for the Coastal Plain and Piedmont regions had significantly different slopes. The results showed important practical differences in percentage of confidence intervals containing the true total volume and in percentage of estimates within a specific proportion of the true total in a simulation study.

KEYWORDS: Mensuration, loblolly pine, practical differences, statistical differences, weighted regression equations.

McClure, Joe P.; Anderson, Jana; Schreuder, Hans T.

A comparison of regional and site-specific volume estimation equations. Res. Pap. SE-264. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1987. 9 pp.

Regression equations for volume by region and site class were examined for loblolly pine. The regressions for the Coastal Plain and Piedmont regions had significantly different slopes. The results showed important practical differences in percentage of confidence intervals containing the true total volume and in percentage of estimates within a specific proportion of the true total in a simulation study.

KEYWORDS: Mensuration, loblolly pine, practical differences, statistical differences, weighted regression equations.