Forest Ecology and Management, 53 (1992) 65–76
Elsevier Science Publishers B.V., Amsterdam

Evaluation of weighted regression and sample size in developing a taper model for loblolly pine

Kenneth L. Cormier[a], Robin M. Reich[a], Raymond L. Czaplewski[b] and William A. Bechtold[c]

[a]*Department of Forest Sciences, Colorado State University, Fort Collins, CO 80523, USA*
[b]*Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO 80526, USA*
[c]*Southeastern Forest Experiment Station, Asheville, NC 28802, USA*

(Accepted 30 September 1991)

ABSTRACT

Cormier, K.L., Reich R.M., Czaplewski R.L. and Bechtold, W.A., 1992. Evaluation of weighted regression and sample size in developing a taper model for loblolly pine. *For. Ecol. Manage.*, 53: 65–76.

A stem profile model, fit using pseudo-likelihood weighted regression, was used to estimate merchantable volume of loblolly pine (*Pinus taeda* L.) in the southeast. The weighted regression increased model fit marginally, but did not substantially increase model performance. In all cases, the unweighted regression models performed as well as the weighted regression models, even for very small sample sizes.

INTRODUCTION

Taper equations can be useful in predicting the volume of an individual tree or a stand of trees. Given the amount of variability that exists among individual tree forms, using smaller sample sizes to evaluate the parameters of a taper equation may increase the risk of making larger errors. The use of weighted regression techniques, however, may reduce the error by properly modeling the variability that exists among individual trees. Disadvantages of using weighted regression are the risk of retransformation errors (Czaplewski and Bruce, 1990) and the relative complexity of using weighted regression on an already complex taper equation.

Biging (1984) developed a taper equation, derived from the Chapman–Richard's function, to describe stem form for second-growth mixed conifers in northern California. The model predicts upper stem diameters as a func-

Correspondence to: K.L. Cormier, Department of Forest Sciences, Colorado State University, Fort Collins, CO 80526, USA.

© 1992 Elsevier Science Publishers B.V. All rights reserved 0378-1127/92/$05.00

tion of relative height and diameter at breast height (dbh). Biging found that his sigmoid equation compared favorably with a segmented polynomial equation developed by Max and Burkhart (1976), which Cao et al. (1980) found to be superior to several other taper models in predicting upper stem diameters, including Kozak et al. (1969), Demaerschalk (1973) and Goulding and Murray (1976). The objectives of this study were to evaluate the effect of weighted regression on the fit and performance of Biging's taper equation, using both large and small data sets.

MATERIALS AND METHODS

Data

The data used for this study were collected on 5350 loblolly pine (*Pinus taeda* L.) trees by the USDA Forest Service, Southeastern Forest Experiment Station, Forest Inventory and Analysis Project (SE FIA), at Asheville, NC. The data were collected over a 20-year period from Florida, Georgia, North Carolina, South Carolina and Virginia. Diameters were measured at the ground, stump height (0.3 m), breast height (1.37 m) and at 4 ft. (1.2 m)-intervals above the stump to a 7 in. (17.5 cm) diameter outside bark (dob). Measurements taken above the 7 in. (17.5 cm) dob were taken at 5 ft. (1.5 m)-intervals to a 4 in. (10 cm) dob.

Forty-two percent of the trees were from randomly selected permanent plots which were measured non-destructively using a McClure mirror caliper and section poles (Cost, 1971). Diameters inside bark (dib) were estimated using prediction equations based on bark thickness at breast height, diameter at breast height (dbh), dob of section ends, and heights of the section ends (Cost, 1978).

The remaining 58% of the trees were sampled from logging operations in the same geographical area. For each state and 10-year inventory cycle, 100 logging sites were selected proportional to the type of wood products (e.g. pulpwood, sawtimber, etc.) being harvested in the inventory unit. At each site, data were taken on as many trees as could be measured in 1 day. Diameters were measured using calipers and heights were measured with steel tapes. Inside bark diameters were measured directly from felled trees using a Swedish bark gauge.

Measurements at ground level were not used because they were suspected of decreasing model performance in the lower bole (McClure and Czaplewski, 1986). A few trees with major defects, broken tops or excessive limbing were excluded because of their unmerchantable or atypical characteristics.

All measured trees were randomly divided into a developmental data set used to estimate the parameters of the taper model, and a validation data set, used for validating the taper model. The development data set contained

40 380 sections from 2742 trees; the validation data set contained 37 828 sections from 2608 trees. The means and SD for dbh and total height were similar for both the development and validation data sets (Table 1).

Four sub-samples of 50 trees each, were randomly sampled from the development data set to assess the importance of using weighted regression in developing taper models using small sample sizes. Means and SD for dbh and total tree height varied among the four sub-samples, with sub-samples one and four having a higher proportion of smaller trees and more variability (Table 1).

Methods

Biging's (1984) sigmoid equation, which has three parameters, was chosen for this study because of its simplicity and the fact that it can easily be inverted to estimate heights. It was also felt that Biging's sigmoidal model would be more sensitive to heterogeneous errors associated with stem profiles.

Biging's sigmoidal model for stem profile is given by

$$(d/D) = b1 + b2\{ln[1-(h/H)^{b3}(1+\exp(-b1/b2)]\} \quad (1)$$

TABLE 1

Descriptive statistics for the development and validation data sets and the four subsets of the development data set. Descriptive statistics for the total data set are also given

Data set	Mean	SD	Min.	Max.	No. of trees	No. of sections
Diameter at breast height (in.)						
Development	10.4	4.331	1.0	33.3	2742	40 380
1st subset	8.3	4.918	1.3	19.3	50	629
2nd subset	10.7	4.818	1.2	20.8	50	776
3rd subset	9.6	4.827	1.5	20.4	50	698
4th subset	8.3	4.916	1.4	18.7	50	678
Validation	10.3	4.325	1.0	38.3	2608	37 828
Total set	10.3	4.328	1.0	38.3	5367	78 208
Total Tree Height (ft.)						
Development	59.8	18.602	9.0	115.0	2742	40 380
1st subset	48.5	23.555	11.0	101.0	50	629
2nd subset	64.6	22.100	18.0	103.0	50	776
3rd subset	59.1	21.161	19.0	111.0	50	698
4th subset	50.7	22.329	11.0	90.0	50	678
Validation	59.6	18.318	9.0	118.7	2608	37 828
Total set	56.7	18.460	9.0	118.7	5367	78 208

where d is diameter at height h, D is diameter at breast height, h is height above ground to top diameter d, H is total tree height, and $b1$,$b2$,$b3$ are regression coefficients. GAUSS™ (1984), a matrix programming language, was used to fit the model to the data using both weighted and unweighted non-linear regression.

Since the variability in diameter is not homogeneous over the range of data, weighted regression might improve the fit. One method commonly used in fitting a weighted regression model is the use of a variance function to approximate the variability associated with the regression model (Fig. 1). Three methods for approximating the variance were suggested by Carrol and Ruppert (1988): (1) squared residuals; (2) absolute residuals; (3) natural log of the absolute residuals. Carrol and Ruppert assumed that the expectations of the squared residuals provided a good approximation of the variance. However, if large outliers exist, then the absolute residuals or the natural log of the absolute residuals can be used. Carrol and Ruppert (1988) suggested using a pseudo-likelihood technique in estimating the variance function. The variance function and estimates of the model parameters were obtained using an iterative procedure: (1) obtain preliminary estimates of coefficients of the taper model using unweighted non-linear regression; (2) based on the plot of residuals, develop a variance function and compute the estimated weights; (3) update the preliminary estimates of the coefficients of the taper model using weighted non-linear regression; (4) repeat steps two and three until estimates of the coefficients for the variance function and the model converge.

To assess the performance of the taper model, estimates of tree volume for the validation data were calculated using diameters and heights predicted from

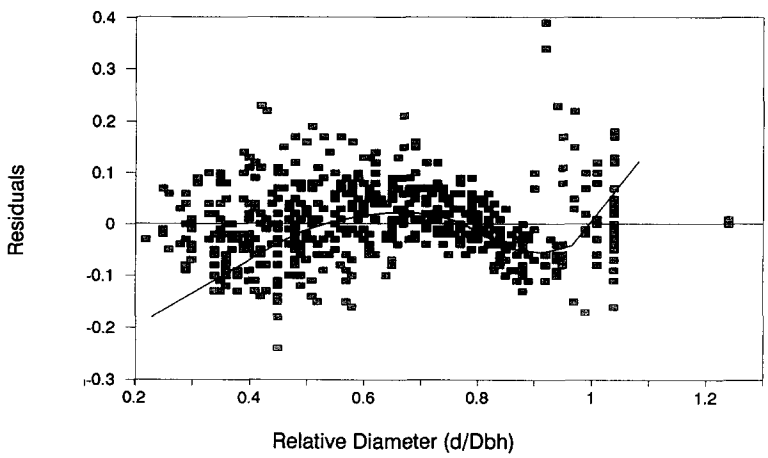

Fig. 1. An example of a variance function calculated from a residual plot using regression techniques.

Biging's taper equation. Predictions of height at any given diameter were derived from Eqn. (1):

$$(h/H)=A-A\{\exp[(d/D)-b1]/b2\}^{1/b3} \quad (2)$$

where $A=1/(1+\exp(\mathrm{bl}/\mathrm{b2}))$. $b1,b2,b3$ are regression coefficients estimated from Eqn. (1).

Board-foot volumes were calculated for sawtimber trees and cubic-foot volumes for pulpwood-sized trees. Sawtimber trees were defined as trees with a dbh of 12 in. or greater and containing at least one 16-ft. sawlog. All other trees were classified as pulpwood. Board-foot volumes were calculated using the Scribner log rule (Avery and Burkhart, 1983) which assumes a 0.25 in. sawkerf. Cubic-foot volumes were calculated using Smalian's formula which assumes a geometric form of paraboloid. In all cases, stump height was assumed to be at 0.5 ft.

To estimate merchantable cubic volume, the length of the stem from the stump to a 3 in. top, estimated using eqn. (2), was divided into 4-ft. sections. Equation (1) was used to predict upper and lower diameters of each 4-ft. section and the volume estimated using Smalian's formula. Cubic volumes of each 4-ft. section were then summed to obtain merchantable cubic foot volume.

Board-foot volumes were estimated for each sawtimber tree assuming an 8-in. merchantable top. Using Eqn. (2), height to an 8-in. top, minus the stump height, was estimated and then divided by 16.25 ft. (16-ft. log plus 0.25 ft. edging) to determine the number of 16-ft. sawlogs in the stem. If there was less than one 16-ft. sawlog, the tree was re-classed as a pulpwood tree and cubic volumes were calculated accordingly. The amount of merchantable sawtimber from the last 16-ft. log to the 8-in. top was calculated by subtracting the height of the last 16-ft. sawlog from the height to the 8-in. top. If the remaining log length was greater than 2 ft., the length was added to the last 16-ft. sawlog, providing it was not the first sawlog, and split as follows:

Log length (ft.)	18	20	22	24	26	28	30
1st Log (ft.)	10	10	12	12	14	14	16
2nd Log (ft.)	8	10	10	12	12	14	14

If the last 16-ft. sawlog was also the first, the remaining sawtimber was counted as a part of topwood. Using Eqn. (1), the diameters at each end of the sawlog were used in estimating board-foot volumes. Topwood volume, the volume between the last sawlog and a 3-in. top, was calculated using the same procedure used for calculating cubic volumes in pulpwood trees.

To assess the behavior of the model, the biases and SD associated with estimating relative diameters and relative heights for the unweighted and weighted models were calculated and compared. The performance of the

weighted and unweighted models was assessed by comparing the percent biases in estimating board-foot and cubic volume, calculated from the validation data.

RESULTS AND DISCUSSION

To compare the performance of the unweighted models with the weighted models, Furnival's index of fit (FIF) was used (Furnival, 1961). Furnival's index of fit enables the comparison of fit of the original unweighted model with the weighted model using computed maximum likelihoods as a criterion for choosing the most appropriate form of the model (Furnival, 1961). Smaller FIF values indicate a better fit to the data, with FIF equaling root mean squared error (RMSE) for an unweighted model. The root mean squared errors between the full development data model and the four subset models were similar within the unweighted and the weighted model groups (Table 2). Although FIF values for the weighted models were consistently smaller than FIF values for the unweighted models, FIF values were not substantially different, indicating that the unweighted models performed nearly as well as the weighted models. The bias and the SD associated with estimat-

TABLE 2

Regression coefficients and relevant fit statistics for the unweighted and weighted models generated from the development data set and the four subsets using Biging's (1984) taper model

Data set	Unweighted model					
	Regression coefficients			RMSE	R^2	FIF
	$b1$	$b2$	$b3$			
Development Data Set						
Unweighted	1.24206	0.29019	0.14970	0.073	0.989	0.073
Weighted	1.21521	0.28513	0.16198	1.013	0.991	0.070
First Subset						
Unweighted	1.20393	0.31651	0.21928	0.083	0.987	0.083
Weighted	1.16097	0.33199	0.27904	1.010	0.990	0.075
Second Subset						
Unweighted	1.19781	0.30048	0.18761	0.076	0.988	0.076
Weighted	1.18693	0.28506	0.17540	1.068	0.992	0.075
Third Subset						
Unweighted	1.25447	0.27788	0.13048	0.075	0.989	0.075
Weighted	1.23238	0.26852	0.13116	1.015	0.991	0.066
Fourth Subset						
Unweighted	1.36363	0.28299	0.10003	0.077	0.989	0.077
Weighted	1.32178	0.28596	0.12171	1.006	0.991	0.072

The root mean square errors (RMSE) for the weighted models are in different units than the RMSE for the unweighted models and should not be compared directly. FIF, Furnival's index of fit.

ing relative diameters (Table 3), and relative heights (Table 4), were similar between the weighted and unweighted models for a given data set.

Volume equations

Estimates of merchantable cubic foot volume of pulpwood-size trees were

TABLE 3

Biases and SE for the unweighted and weighted models generated from the development data set and the four subsets using eqn. (1): predicting relative diameter across relative height

Relative height	Bias		SE	
	Unweighted	Weighted	Unweighted	Weighted
Development data set				
0.0–0.1	0.015	0.021	0.103	0.105
0.1–0.2	−0.043	−0.042	0.067	0.033
0.2–0.4	0.001	0.001	0.054	0.054
0.4–0.6	0.015	0.011	0.035	0.065
0.6–0.8	−0.008	−0.013	0.074	0.075
0.8–1.0	−0.011	−0.013	0.046	0.048
1st Subset				
0.0–0.1	−0.008	−0.005	0.104	0.106
0.1–0.2	−0.070	−0.074	0.088	0.090
0.2–0.4	−0.024	−0.029	0.059	0.031
0.4–0.6	−0.005	−0.009	0.064	0.064
0.6–0.8	−0.020	−0.022	0.076	0.077
0.8–1.0	−0.013	−0.013	0.047	0.047
2nd Subset				
0.0–0.1	0.020	0.033	0.105	0.109
0.1–0.2	−0.042	−0.032	0.067	0.061
0.2–0.4	0.001	0.008	0.054	0.054
0.4–0.6	0.016	0.017	0.066	0.066
0.6–0.8	−0.005	−0.008	0.074	0.075
0.8–1.0	−0.009	−0.012	0.045	0.046
3rd Subset				
0.0–0.1	0.022	−0.034	0.104	0.109
0.1–0.2	−0.037	−0.030	0.064	0.060
0.2–0.4	0.004	0.008	0.054	0.052
0.4–0.6	0.014	0.015	0.066	0.066
0.6–0.8	−0.017	−0.008	0.075	0.075
0.8–1.0	−0.013	−0.012	0.045	0.049
4th Subset				
0.0–0.1	−0.022	−0.022	0.103	0.103
0.1–0.2	−0.072	−0.075	0.089	0.092
0.2–0.4	−0.025	−0.029	0.060	0.062
0.4–0.6	−0.010	−0.014	0.064	0.065
0.6–0.8	−0.031	−0.034	0.080	0.081
0.8–1.0	−0.020	−0.021	0.052	0.053

TABLE 4

Biases and SE for the unweighted and weighted models generated from the development data set and the four subsets using eqn. (2): predicting relative height across relative diameter

Relative diameter	Bias		SE	
	Unweighted	Weighted	Unweighted	Weighted
Development data set				
0.0–0.2	−0.007	−0.055	0.007	0.006
0.2–0.4	−0.051	−0.055	0.006	0.006
0.4–0.6	−0.003	−0.008	0.006	0.006
0.6–0.8	0.009	0.007	0.008	0.008
0.8–1.0	−0.012	−0.011	0.005	0.005
1st Subset				
0.0–0.2	−0.007	−0.094	0.001	0.012
0.2–0.4	−0.056	−0.055	0.006	0.006
0.4–0.6	−0.021	−0.023	0.006	0.006
0.6–0.8	−0.025	−0.031	0.009	0.009
0.8–1.0	−0.044	−0.049	0.007	0.007
2nd Subset				
0.0–0.2	−0.007	−0.045	0.007	0.005
0.2–0.4	−0.048	−0.052	0.005	0.006
0.4–0.6	0.001	−0.001	0.006	0.006
0.6–0.8	0.010	0.018	0.008	0.008
0.8–1.0	−0.012	0.001	0.005	0.005
3rd Subset				
0.0–0.2	−0.007	−0.043	0.001	0.005
0.2–0.4	−0.055	−0.059	0.006	0.007
0.4–0.6	−0.006	−0.007	0.006	0.006
0.6–0.8	0.012	0.018	0.008	0.008
0.8–1.0	0.007	0.003	0.005	0.005
4th Subset				
0.0–0.1	−0.008	−0.095	0.001	0.013
0.2–0.4	−0.067	−0.069	0.008	0.008
0.4–0.6	−0.031	−0.035	0.007	0.007
0.6–0.8	−0.028	−0.034	0.009	0.009
0.8–1.0	−0.046	−0.050	0.007	0.007

consistently overpredicted across all five data sets. These overpredictions are believed to be caused by retransformation bias. Czaplewski and Bruce (1990) found retransformation bias affects predictions for volume in profile models that are fit to d/D.

There was little difference between the weighted and unweighted subset models. Both the weighted and unweighted models for two of the four small sample subsets (2 and 3) were in close agreement with the development data set, whereas the models for the other two subsets (1 and 4) were substantially different (Fig. 2).

Using the model developed from the original development data set, board-

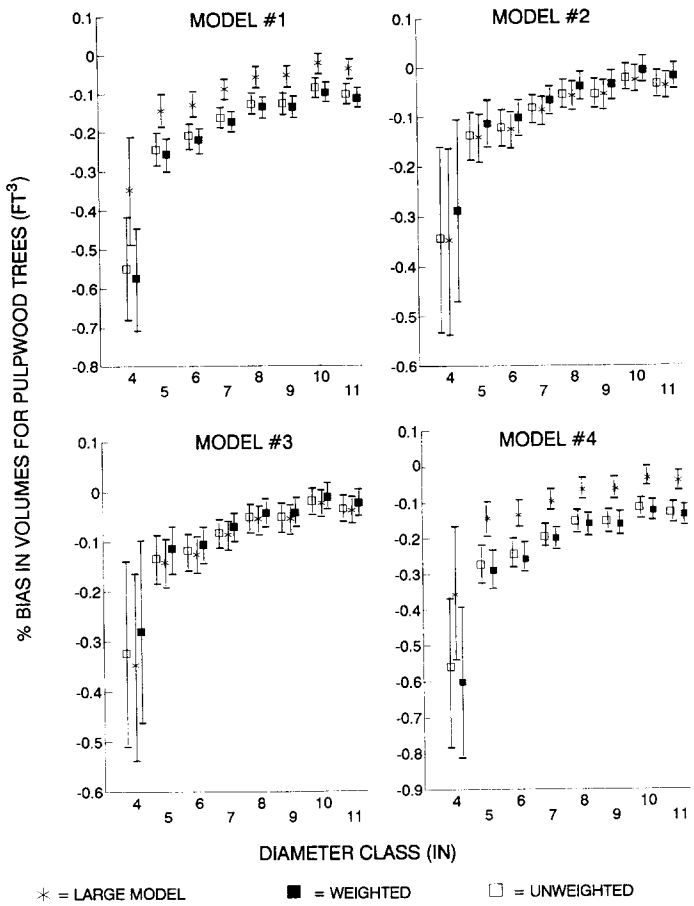

Fig. 2. Plot of the percent bias in pulpwood volume by diameter class for the large data set and each of the four subsets with small sample sizes. The large, unweighted data set was used to compare the accuracy of the weighted and unweighted models for each subset. As the scales are different for each plot, the large data set also provides a standard for comparison. Negative values represent over-predictions; positive values represent under-predictions. The error bars represent 95% confidence intervals.

foot volume estimates underpredicted volume in the smaller diameter classes and overpredicted in the larger diameter classes. Again there appeared to be little difference between the weighted and unweighted models for each subset (Fig. 3). Both the weighted and unweighted models for subsets two and three agreed with the development data set, whereas the models for subsets one and four overpredicted volumes across all diameter classes (Fig. 3).

Topwood volume estimates were consistently overestimated across all diameter classes using the model developed with the original development data set. Once again, there was little difference between the weighted and un-

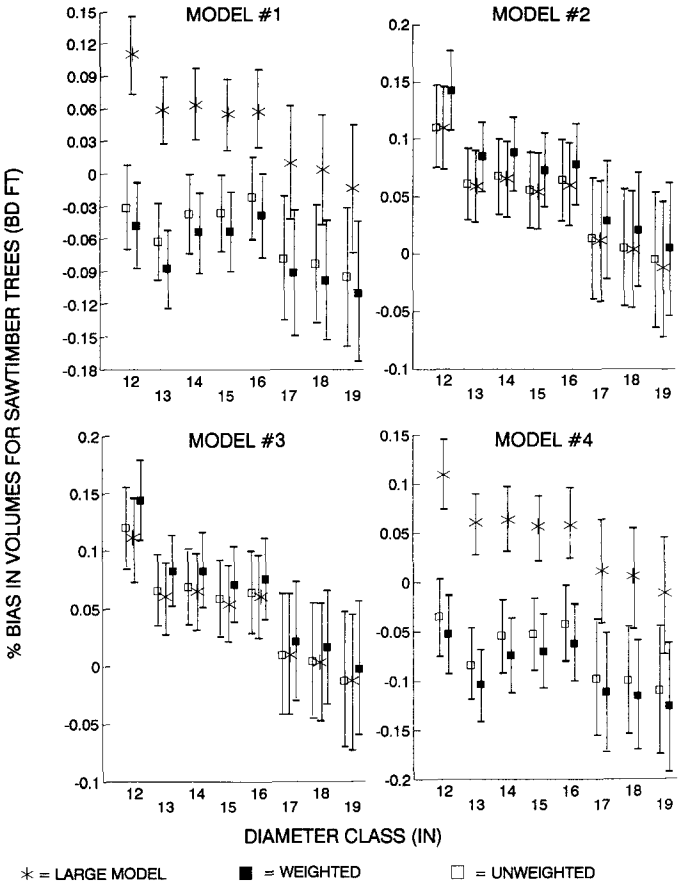

Fig. 3. Plot of the percent bias in sawtimber volumes by diameter class for the large data set and each of the four subsets with small sample sizes. The large, unweighted data set was used to compare the accuracy of the weighted and unweighted models for each subset. As the scales are different for each plot, the large data set also provides a standard for comparison. Negative values represent over-predictions; positive values represent under-predictions. The error bars represent 95% confidence intervals.

weighted models for each subset (Fig. 4). Again the models for two subsets agreed with the development data set, whereas the models for the other two subsets overpredicted volume at higher levels (Fig. 4). As the errors in estimating topwood volume are negatively correlated with errors in estimating board-foot volumes for the same trees, it was not surprising to find topwood volumes being overpredicted when board-foot volumes were underestimated.

In all three volume estimations, both the weighted and unweighted models for subsets one and four did not give adequate estimations of volume. A possible explanation for this is that subsets one and four did not adequately describe the population because of sampling errors. Mean dbh and range of dbh

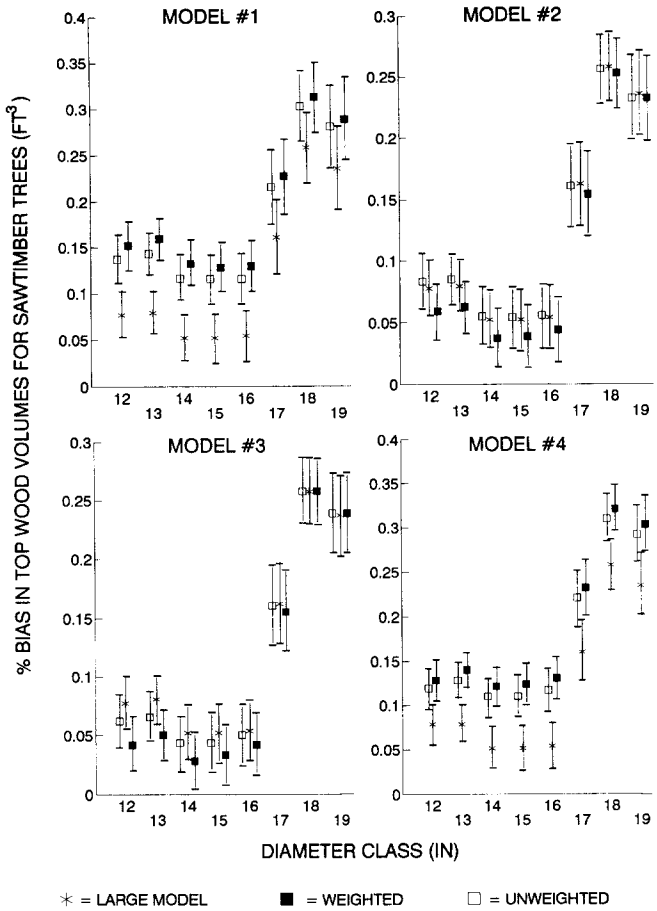

Fig. 4. Plot of the percent bias in sawtimber topwood volumes by diameter class for the large data set and each of the four subsets with small sample sizes. The large, unweighted data set was used to compare the accuracy of the weighted and unweighted models for each subset. As the scales are different for each plot, the large data set also provides a standard for comparison. Negative values represent over-predictions; positive values represent under-predictions. The error bars represent 95% confidence intervals.

for subsets one and four were smaller than mean dbh and range of dbh for the other data sets, indicating that subsets one and four had a higher concentration of smaller diameter trees, and fewer large diameter trees. This would cause an inherent bias in the models developed from subsets one and four. Table 3 indicates that the models developed from subsets one and four consistently overpredicted relative diameters across the entire tree length, whereas the models developed from the other data sets overpredicted relative diameters at the base and the top of the tree.

CONCLUSIONS

Although the pseudo-likelihood approach has been shown to be a good method for fitting weighted regression models (Carrol and Ruppert, 1988), we found that it did not substantially increase model performance for Biging's (1984) stem profile model. The weighted models had smaller errors associated with fitting the data but the errors were not substantially smaller than the unweighted models.

Using Biging's (1984) stem profile model, we observed that sub-samples of 50 trees estimated volumes as well as the large sample of 2742 trees in two out of four cases. The two sub-samples that failed to estimate volumes satisfactorily did so because they did not adequately represent the population. The weighted regression technique did not reduce the risk of errors associated with using smaller sample sizes with respect to Biging's taper equation.

REFERENCES

Avery, T.E. and Burkhart, H.E., 1983. Forest Measurements. McGraw-Hill, New York, 331 pp.

Biging, G.S., 1984. Taper equations for second-growth mixed conifers of Northern California. For. Sci., 30: 1103–1117.

Cao, Q.V., Burkhart, H.E. and Max, T.A., 1980. Evaluation of two methods for cubic-volume prediction of loblolly pine to any merchantable limit. For. Sci., 26: 71–80.

Carroll, R.J. and Ruppert, D., 1988. Transformation and weighting in regression. Chapman and Hall, New York, pp. 69–84.

Cost, N.D., 1971. Accuracy of Standing-Tree Volume Estimates Based on McClure Mirror Caliper Measurements. USDA Forest Service, Southeast. For. Exp. Stn. Res. Note SE-152, 4 pp.

Cost, N.D., 1978. Multiresource Inventories – A Technique for Measuring Volumes in Standing Trees. USDA Forest Service, Southeast. For. Exp. Stn. Res. Paper SE-196, 18 pp.

Czaplewski, R.L. and Bruce, D., 1990. Retransformation bias in a stem profile model. Can. J. For. Res., 20: 1623–1630.

Demaerschalk, J.P., 1973. Integrated systems for the estimation of tree taper and volume. Can. J. For. Res., 3: 90–94.

Furnival, G.M., 1961. An index for comparing equations used in constructing volume tables. For. Sci., 7: 337–341.

Gauss, 1988. Aptech Systems, 26250-196th PL.SE. Kent, WA.

Goulding, C. J. and Murray, J.C., 1976. Polynomial taper equations that are compatible with tree volume equations. N. Z. J. For. Sci., 5: 313–322.

Kozak, A., Munro, D.D. and Smith, J.H.G., 1969. Taper functions and their application in forest inventory. For. Chron., 45: 278–283.

Max, T.A. and Burkhart, H.E., 1976. Segmented polynomial regression applied to taper equations. For. Sci., 22: 283–289.

McClure, J.P. and Czaplewski, R.L., 1986. Compatible taper equation for loblolly pine. Can. J. For. Res., 16: 1272–1277.