

TABLE OF CONTENTS

PART I: TUTORIAL PAPERS

Combining the Error of Sample Plots and Biomass Regressions

- Error of forest inventory estimates: its main components 1
Tiberius Cunia
- An optimization model to calculate the number of sample trees and plots 15
Tiberius Cunia

Error of Biomass Regressions

- Construction of tree biomass tables by linear regression techniques 27
Tiberius Cunia
- Use of dummy variables techniques in the estimation of biomass regressions 37
Tiberius Cunia
- On the error of tree biomass regressions: trees selected by cluster sampling and double
sampling 49
Tiberius Cunia

Error of Sample Plots

- On the error of forest inventory estimates: stratified sampling and double sampling for
stratification 63
Tiberius Cunia
- On the error of forest inventory estimates: two-stage sampling of plots 71
Tiberius Cunia
- On the error of forest inventory estimates: double sampling with regression 79
Tiberius Cunia
- On the error of forest inventory estimates: Continuous Forest Inventory without SPR 89
Tiberius Cunia
- On the error of forest inventory estimates: Continuous Forest Inventory with SPR 99
Tiberius Cunia

PART II: RESEARCH PAPERS

Biomass Regressions and Measurement Error

- An optimization model for subsampling trees for biomass measurement 109
Tiberius Cunia
- Estimating sample tree biomass by subsampling: some empirical results 119
R. D. Briggs, T. Cunia, E. H. White, and H. W. Yawney
- Unbiased estimation of total tree weight by three-stage sampling with probability
proportional to size 129
Harry T. Valentine, Timothy G. Gregoire, and George M. Furnival
- Measurement errors in forest biomass estimation 133
Daniel Auclair

Biomass of Forest Understory Vegetation

- Biomass-dimension relationships of understory vegetation in relation to site and stand
age 141
Paul B. Alaback

TABLE OF CONTENTS

Biomass estimates for nontimber vegetation in the Tanana River Basin of Interior Alaska	149
Bert Mead, John Yarie, and David Herman	

Biomass Functions in the Eastern United States: Regression Models and Application to Timber Inventories

A summary of equations for predicting biomass of planted southern pines	157
V. C. Baldwin, Jr.	
Summary of biomass equations available for softwood and hardwood species in the southern United States	173
Alexander Clark III	
Methods for estimating the forest biomass in Tennessee Valley Region	189
J. Daniel Thomas and Robert T. Brooks, Jr.	
Areas of biomass research ¹	193
Boris Zeide	

Biomass Studies Outside the United States

Prediction error in tree biomass regression functions for western Canada	199
T. Singh	
Forest biomass studies in France	209
Daniel Auclair	
Biomass studies in Europe - an overview	213
Dieter R. Pelz	
Subsampling trees for biomass	225
C. Kleinn and D. R. Pelz	
Simple biomass regression equations for subtropical dry forest species	229
Joseph D. Kasile	

Use of Simulation Techniques to Evaluate the Validity of Biomass Regression Functions

Evaluating errors of tree biomass regressions by simulation	235
Tiberius Cunia	
Estimation of tree biomass tables by cluster sampling: results of a simulation study	243
Andrew J. Gillespie and Tiberius Cunia	
Error of biomass regressions: sample trees selected by stratified sampling	253
Alexandros Arabatzis and Tiberius Cunia	
Error of biomass regressions: sample trees selected by double sampling	269
John Michelakackis and Tiberius Cunia	
Using simulation to evaluate volume equation error and sampling error in a two-phase design	287
David C. Chojnacky	
High order regression models for regional volume equations	295
Joe P. McClure and Raymond L. Czaplewski	

¹Contributed paper, not presented at the workshop.

SELECTED BY DOUBLE SAMPLING

John Michelakackis and Tiberius Cunia

Graduate student and Professor of Statistics and Operations Research, respectively, State University of New York, College of Environmental Science and Forestry, Syracuse, NY, 13210

Sometimes it is useful to construct biomass tables by a double sampling technique whereby (i) the first phase sample provides a linear regression $\hat{y} = r_1(d, h)$ of tree biomass on tree diameter d and height h , (ii) the second phase sample provides a linear regression $\hat{h} = r_2(d)$ of tree height h on diameter d and (iii) the regression of biomass on diameter alone is defined as $\hat{y} = r(d) = r_1(d, r_2(d))$. By simulation techniques, sets of one hundred samples were drawn from a known tree population by a variety of double sampling (two-phase, two-stage) techniques, one set for each sampling method. Using a variety of procedures, the regression functions of the form $\hat{y} = r(d)$ were first estimated from the data of each sample and then applied to the parent population to estimate the known value of the average biomass per acre. By analyzing the probabilistic behavior of the estimates, inferences were then made about the bias, precision and sample-based estimates of the precision for each combination of sampling method and estimation procedure.

Introduction

In two previous papers, Cunia (1982) and Cunia and Michelakackis (1983c) proposed a method to use data from trees selected by a double (two-phase) sampling technique to estimate biomass regression functions and their error. The trees of the first phase are measured for biomass y , diameter d and height h and their data are used to estimate a linear regression function $\hat{y} = r_1(d, h)$ of biomass on diameter and height. The trees of the second phase are measured for d and h alone, and their data are used to estimate a linear regression function $\hat{h} = r_2(d)$ of height on diameter. The regression function of biomass on diameter alone is estimated by the function $\hat{y} = r(d) = r_1(d, \hat{h}) = r_1(d, r_2(d))$. It is assumed that the error of $r_1(d, h)$ and $r_2(d)$ can be properly evaluated and, thus, their error can be combined to estimate the error of $\hat{y} = r(d)$ by a technique described in their papers.

To test this method, Michelakackis and Cunia (1985) used simulated sampling. In the first phase of the double sampling design above, they have randomly selected, by computer, 15 percent of the trees from 30 clusters (plots of fixed area) also randomly selected from a known forest tree population. These trees were measured for biomass, diameter and height. In the second

phase, 11.7 percent of the trees from 200 randomly selected plots were measured for diameter and height alone. Using the data from one individual two-phase sample, they calculated (i) the regressions $\hat{y} = r_1(d, h)$, $\hat{h} = r_2(d)$ and $\hat{y} = r(d)$ and (ii) the estimates z of the average biomass per acre μ for the given population, $(z - \mu)$ of the bias of the sampling method and estimation procedure and V of the variance of z , using the basic assumptions of the statistical models they worked with in the estimation of the regression functions above.

This sampling and estimation procedure was repeated 100 times, and 100 estimates z and V were thus obtained, together with the average estimates \bar{z} of μ , $(\bar{z} - \mu)$ of the bias of z and \bar{V} of the variance of z . Because the 100 values z were viewed as generated by the same, independently performed random process, they calculated also another estimate of the variance of z , this time unbiased, namely $S_{zz} = \sum (z - \bar{z})^2 / 99$, where \sum means summation over the 100 simulated samples.

For the calculation of the regression functions $\hat{y} = r_1(d, h)$ Michelakackis and Cunia (1985) have used four estimation approaches and, for each estimation approach, several regression equations. The four estimation approaches consisted of two ordinary least (OLS) and weighted least squares (OWLS) and two modified least (MLS) and weighted least squares (MWLS) methods. As the conditional variance of h given d is approximately homogeneous (at least for our tree population) only the least squares OLS and MLS approaches were used in the estimation of $\hat{h} = r_2(d)$. These four approaches will also be used in this study. For a more detailed description of these approaches the reader is referred to the above mentioned paper or additional papers by Cunia (1979, 1981) and Briggs and Cunia (1982). For the purpose of the present study, it suffices to say that (i) the ordinary least squares techniques are applied to individual tree data and they ignore the cluster effect, if any, and (ii) the modified least squares are applied to cluster variables and they take the cluster effect into account.

The analysis of their simulated sampling results showed that (i) for all four estimation approaches, the best of the first and second phase regression functions are $\hat{y} = \alpha_1 + \alpha_2 d^2 h$ and $\hat{h} = \gamma_1 + \gamma_2 d + \gamma_3 d^2$ respectively, (ii) using these two regression equations, four double sampling regression functions $\hat{y} = r(d)$ were defined and applied to the tree population, one for each estimation approach, (iii) all four corresponding estimates z were either unbiased or with a bias so small that in all cases it was not significantly different from zero, (iv) the best estimator z of μ was obtained by the OWLS method (with the OLS estimator following closely) but the statistic V grossly underestimated the variance of z and, finally (v) the best estimator of the variance of z was obtained by the MLS method, but the corresponding estimator z was not as precise as the estimator z obtained by the OWLS method.

The conclusions they reached are strictly valid for the population they worked with and the sampling and estimation procedures they have specifically considered. But they are also indicative of what one should expect in real life. Because the population was constructed from empirical data so as to imitate what really happens in the real world, it is probable that only the strength, not the type of conclusions reached is affected. With respect to sampling and estimation procedures, however, further research may be needed.

It is the objective of the present study to extend the scope of the Michelakackis and Cunia (1985) study and investigate the generality of their conclusions when (i) we vary the number of sample clusters and the number of trees selected from these clusters, (ii) we select the trees from the sample clusters with unequal probability, (iii) we select a fixed rather than a percentage of trees from the sample clusters and (iv) the sampling is done with replacement. The selection of the sample clusters themselves will still be done by the same sampling procedure; single random sampling without replacement.

The population of trees from which the simulated samples will be selected remains the same. The method by which it was constructed is described in detail in a series of papers by Cunia and Michelakackis (1983a, 1984a,b and Cunia, Michelakackis and Lee (1984) and summarized by Michelakackis and Cunia (1985). The interested reader is referred to the above-mentioned papers for more details.

Sampling Method

The basic sampling design considered here has been defined as a double sampling or, more specifically as a two-phase, two-stage sampling design. Except for the sample size and the method of selecting the sample trees (of the second stage) from the sample plots (of the first stage), the two-stage sampling method used in each phase is identical to that described in the introductory section or, in more detail, in papers by Michelakackis and Cunia (1985) and Cunia (1986).

In each of the two phases we have used the following two-stage basic design. In the first stage m clusters (plots) are selected by single random sampling without replacement. The clusters selected in the first stage are further subsampled in the second stage; sample trees are selected from each sample cluster by one of seven basic subsampling procedures. The sample size is controlled by the number m of sample clusters and number of sample trees per cluster. The first phase trees are measured for their diameter d , height h and biomass y . The second phase trees, however, are only measured for d and h ; they are no longer measured for y .

The seven subsampling procedures are better described in terms of the following attributes:

whether the trees are selected with or without replacement, whether a fixed number or a fixed percentage of trees are selected from the sample clusters, whether the probability of tree selection is equal or proportional to h (height), d (diameter), d^2 (basal area) and d^2h (approximate volume) and whether a given number of sample trees is selected from a small or a large number m of sample clusters. For the first phase, the number of clusters we have sampled is $m = 1, 2, 5, 10, 15, 20, 30$ and 50 , the fixed number of trees per clusters we have used is $r = 1, 2, 5, 10, 15, 20$ and 30 and the fixed percentage is $p = 5, 10, 15, 30, 40, 60$ and 100 . For the second phase we have used the values $m = 50, 100, 150, 200, 300$ and 400 , $r = 1, 2, 3$ and 4 , and $p = 2.93, 5.86, 8.79$ and 11.72 . It may be of interest to mention here the fact that when $p = 2.93, 5.86, 8.79$ and 11.72 , the expected number of sample trees per cluster is $1, 2, 3$, and 4 respectively. Not all possible combinations of the attributes above were used in the present study. For example, because of the type of population we have constructed (with trees distributed in one-fifth acre plots) it was not possible to select trees with unequal probability and without replacement.

More specifically, we shall define the following seven subsampling procedures, which for convenience will be known here as the seven basic sampling methods used in each phase of the double sampling design. In method 1, a fixed percentage of trees is selected from each sample cluster, with equal probability and without replacement. When the fixed percentage is replaced by a fixed number of trees, we obtain basic method 2, if the tree selection is done without replacement, or basic method 3, if the selection is done with replacement. By making the probability of tree selection of method 3 proportional to h , d , d^2 and d^2h , we obtain the corresponding basic methods 4, 5, 6 and 7.

Note that the probability of selection refers to the selection of trees from within a sample cluster and may, or may not be the same as the probability of tree selection from the entire population. It is known that the tree size is related to the number of trees contained in a plot of fixed area. Large trees require more space to grow than small trees and, thus, the trees from plots with large number of trees must be relatively small in size, on the average. Consequently, a selection with equal probability of (i) plots of fixed area and (ii) fixed number of trees from each selected plot will not necessarily result in a sample of trees selected with equal probability from the overall population. Large trees, selected mostly from plots with small number of trees are much more likely to be included in the sample than small trees.

When the simulation process was applied and the simulated samples were produced, several problems of the practical order were encountered. For example, multiplication of the fixed percentage p by the number of trees, say n in a given sample cluster does not generally result in an

integer number (np) of trees to be sampled. To decide whether to select or not to select an additional sample tree from the given plot (corresponding to the fractional part of np) we had to devise a Monte Carlo procedure. Furthermore, when the trees are selected with replacement, the same tree may have to be included in the sample more than once. In order to have "different" trees in the sample, we have used the diameter, height and species of the tree selected more than once and generated a new value for its biomass, using the same procedure we have previously used to construct the population. Finally, to reduce the amount of simulation work required, we have selected the sample clusters and the sample trees within these clusters in a nested fashion. These problems were discussed elsewhere and the interested reader is referred to Cunia (1986) for more details.

Because of the large number of samples generated by the 100 simulation runs, the sample data were stored in two sets of seven tapes each; one set for each of the two phases and one tape for each of the seven basic sampling methods. Pairing a sample from one set with a sample from the other set defines a two-phase, two-stage sample from whose data biomass regression functions are being calculated. There is an enormous number of such pairs that one can define, resulting in an astronomical number of two-phase, two-stage samples. To reduce this number by a factor of 10000, we decided to pair samples only if they come from the same simulation run. To further reduce the number of double samples, we have also decided to put aside, at least for the time being, the data from six tapes of phase 2, those containing the samples obtained by the basic sampling methods 2, 3, ..., 7. This implies that, in our present study, we have paired samples obtained by all seven basic subsampling procedures of phase 1 with samples obtained only with the first sampling method of phase 2. Further reductions were also made arbitrarily; they are not mentioned here.

Estimation Procedures

To estimate the average biomass per acre of our tree population, when data from a two-phase, two-stage sample are given, one may use a wide variety of estimation procedures. Consisting of a fixed set of calculation rules, an estimation procedure is defined in our study in terms of (i) a least squares estimation approach applied to (ii) a linear regression equation of a given form using (iii) a given set of sample tree data. As this terminology is specific to our study, let us define in more detail the elements of the various estimation procedures used here.

To estimate the coefficients of the various regression functions, we have used the four least squares estimation approaches briefly described in the introductory section as the ordinary least (OLS) and weighted least squares (OWLS) and the modified least (MLS) and weighted least squares (MWLS) methods. We have used all these four approaches to estimate the coefficients of $\hat{y} = r_1(d, h)$, the regression function of tree biomass

on diameter and height and the coefficients of $\hat{y} = r_3(d)$, the regression function of biomass on diameter alone, using only the data from the first phase sample. However, to estimate the coefficients of $\hat{h} = r_2(d)$, the regression function of tree height on diameter from the data of the second phase sample, we have used only the OLS and MLS approaches; the conditional variance of tree height given diameter is homogeneous.

Michelakackis and Cunia (1985) have found that the regression functions $\hat{y} = r_1(d, h) = \alpha_1 + \alpha_2 d^2 h$ and $\hat{y} = r_3(d) = \beta_1 + \beta_2 d^2$ of the first phase and $\hat{h} = r_2(d) = \gamma_1 + \gamma_2 d + \gamma_3 d^2$ of the second phase were the best from among the several alternate regression functions that they have considered. This was not surprising; it is consistent with the procedures used to construct the population of trees. To simplify our study, we have decided to use only these functions and, thus, work with the double sampling biomass regression functions defined as

$$\begin{aligned}\hat{y} = r(d) &= r_1(d, \hat{h}) = a_1 + a_2 d^2 \hat{h} \\ &= a_1 + a_2 d^2 (c_1 + c_2 d + c_3 d^2) \\ &= b_1 + b_2 d^2 + b_3 d^3 + b_4 d^4\end{aligned}$$

where $a_1, a_2, c_1, c_2,$ and c_3 are the estimates

of $\alpha_1, \alpha_2, \gamma_1, \gamma_2$ and γ_3 respectively and

$$b_1 = a_1, b_2 = a_2 c_1, b_3 = a_2 c_2 \text{ and } b_4 = a_2 c_3$$

To calculate the estimate $[S_{bb}]$ of the covariance matrix of the vector $[b]^T = [b_1 b_2 b_3 b_4]$ of regression coefficients, when the vectors $[a]^T = [a_1 a_2]$ and $[c]^T = [c_1 c_2 c_3]$ are given together with the estimates $[S_{aa}]$ and $[S_{cc}]$ of their covariance matrices, we have used the procedures described by Cunia (1982) and illustrated by Cunia and Michelakackis (1983c). Because these procedures have been streamlined and made better for computer applications, and this new streamlined version has not been given before, let us describe it below, in the more general form as used by our computer program.

We start by expressing the estimators of the more general regression functions $\hat{y} = r_1(d, h)$ and $\hat{h} = r_2(d)$ as

$$\begin{aligned}\hat{y} &= a_1 + a_2 d^2 h + a_3 d + a_4 h + a_5 d h + a_6 d^2 \\ \text{and} \quad \hat{h} &= c_1 + c_2 d + c_3 d^2\end{aligned}$$

Let us also express the covariance matrices $[S_{aa}]$ and $[S_{cc}]$ of $[a]$ and $[c]$ respectively in their more explicit form

$$[S_{aa}] = \begin{bmatrix} S_{a_1 a_1} & S_{a_1 a_2} & \dots & S_{a_1 a_6} \\ S_{a_1 a_2} & S_{a_2 a_2} & \dots & S_{a_2 a_6} \\ \vdots & \vdots & \ddots & \vdots \\ S_{a_1 a_6} & S_{a_2 a_6} & \dots & S_{a_6 a_6} \end{bmatrix}$$

and

$$[S_{cc}] = \begin{bmatrix} s_{c_1 c_1} & s_{c_1 c_2} & s_{c_1 c_3} \\ s_{c_1 c_2} & s_{c_2 c_2} & s_{c_2 c_3} \\ s_{c_1 c_3} & s_{c_2 c_3} & s_{c_3 c_3} \end{bmatrix}$$

Some of the coefficients a and c may be made equal to zero (when they are not significantly different than zero) and, thus, the corresponding rows and columns of $[S_{aa}]$ and $[S_{cc}]$ will also be made equal to zero. For example, when $y = a_1 + a_2 d^2$ and $h = c_1 + c_2 d + c_3 d^2$, then

$$[a]' = [a_1 \quad a_2 \quad 0 \quad 0 \quad 0 \quad 0]$$

$$[c]' = [c_1 \quad c_2 \quad c_3]$$

$$[S_{aa}] = \begin{bmatrix} s_{a_1 a_1} & s_{a_1 a_2} & 0 & \dots & 0 \\ s_{a_1 a_2} & s_{a_2 a_2} & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

and $[S_{cc}]$ has the same form as that shown above.

Returning now to the more general case, it can be shown by lengthy but straightforward algebraic calculations, that the regression function $\hat{y} = r_1(d, h) = r(d)$ can be written as

$$\hat{y} = [b]'[x] = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5$$

where $x_1 = 1$, $x_2 = d$, $x_3 = d^2$, $x_4 = d^3$ and $x_5 =$

d^4 , and $b_1 = (a_1 + a_4 c_1)$, $b_2 = (a_3 + a_4 c_2 +$

$a_5 c_1)$, $b_3 = (a_2 c_1 + a_4 c_3 + a_5 c_2 + a_6)$, $b_4 = (a_2 c_2$

$+ a_5 c_3)$ and $b_5 = a_2 c_3$. In matrix notation we can

write

$$[b] = [C][a] = [A][c_e]$$

where

$$[C] = \begin{bmatrix} 1 & 0 & 0 & c_1 & 0 & 0 \\ 0 & 0 & 1 & c_2 & c_1 & 0 \\ 0 & c_1 & 0 & c_3 & c_2 & 1 \\ 0 & c_2 & 0 & 0 & c_3 & 0 \\ 0 & c_3 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$[A] = \begin{bmatrix} a_1 & a_4 & 0 & 0 \\ a_3 & a_5 & a_4 & 0 \\ a_6 & a_2 & a_5 & a_4 \\ 0 & 0 & a_2 & a_5 \\ 0 & 0 & 0 & a_2 \end{bmatrix}$$

$$[a]' = [a_1 \quad a_2 \quad a_3 \quad a_4 \quad a_5 \quad a_6]$$

$$[c_e]' = [1 \quad c_1 \quad c_2 \quad c_3]$$

Finally, it can be shown that the covariance matrix of $[b]$ can be estimated by the (approximate) formula

$$[S_{bb}] = [C][S_{aa}][C]' + [A][S_{c_e c_e}][A]'$$

where $[S_{c_e c_e}]$ is the estimate of the covariance matrix of the expanded vector $[c_e]$, that is

$$[S_{c_e c_e}] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & s_{c_1 c_1} & s_{c_1 c_2} & s_{c_1 c_3} \\ 0 & s_{c_1 c_2} & s_{c_2 c_2} & s_{c_2 c_3} \\ 0 & s_{c_1 c_3} & s_{c_2 c_3} & s_{c_3 c_3} \end{bmatrix}$$

When some of the regression coefficients a and c are made equal to zero, it suffices to write zero whenever these coefficients occur in the formulae above and substitute zero for all their variance or covariance terms. This would be the case in the present study where $a_3 = a_4 = a_5 = a_6 = 0$. Then it can be shown that

$$[b]' = [a_1 \quad 0 \quad a_2 c_1 \quad a_2 c_2 \quad a_2 c_3]$$

and

$$[S_{bb}] = [C][S_{aa}][C]' + [A][S_{c_e c_e}][A]'$$

$$= \begin{bmatrix} s_{a_1 a_1} & 0 & c_1 s_{a_1 a_2} & c_2 s_{a_1 a_2} & c_3 s_{a_1 a_2} \\ 0 & 0 & 0 & 0 & 0 \\ c_1 s_{a_1 a_2} & 0 & c_1^2 s_{a_2 a_2} & c_1 c_2 s_{a_2 a_2} & c_1 c_3 s_{a_2 a_2} \\ c_2 s_{a_1 a_2} & 0 & c_1 c_2 s_{a_2 a_2} & c_2^2 s_{a_2 a_2} & c_2 c_3 s_{a_2 a_2} \\ c_3 s_{a_1 a_2} & 0 & c_1 c_3 s_{a_2 a_2} & c_2 c_3 s_{a_2 a_2} & c_3^2 s_{a_2 a_2} \end{bmatrix}$$

$$+ a_2^2 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & s_{c_1 c_1} & s_{c_1 c_2} & s_{c_1 c_3} \\ 0 & 0 & s_{c_1 c_2} & s_{c_2 c_2} & s_{c_2 c_3} \\ 0 & 0 & s_{c_1 c_3} & s_{c_2 c_3} & s_{c_3 c_3} \end{bmatrix}$$

Except for the fact that b_2 (the coefficient of d) is not included in the regression function and the other regression coefficients b_3 , b_4 , and b_5 are renumbered as b_2 , b_3 and b_4 respectively, the regression function $\hat{y} = r(d)$ is identical to the double sampling regression function defined before as

$$\hat{y} = b_1 + b_2 d^2 + b_3 d^3 + b_4 d^4$$

Let us denote a linear regression function of tree biomass on diameter and possibly height by the general expression

$$y = [b]'[x] = b_1x_1 + b_2x_2 + \dots + b_mx_m$$

where $x_1 = 1$ and x_2, x_3, \dots, x_m are functions of diameter d and possibly height h . Then, one can calculate this regression from different sets of sample data and, subsequently calculate the associated estimator of μ , and its error, by the formulae

$$z = [b]'[\mu_x] = \text{estimator of } \mu, \text{ and}$$

$$V = [\mu_x]'[S_{bb}][\mu_x] = \text{estimator of the variance of } z,$$

where $[\mu_x]$ is the vector of the population parameters $\mu_1, \mu_2, \dots, \mu_m$ defined as the "means per acre" of the sums of variables x_1, x_2, \dots, x_m respectively, expressed on a per acre basis. For more details on these formulae the reader should refer to Cunia and Michelakackis (1983c) and Cunia (1986).

For each combined sample of the first and second phase, that is, for each double sample, we shall consider three estimators, each estimator being based on a specific set of sample data. The first, to be known here as the first single sampling estimator, is based on the regression function $\hat{y} = r_1(d, h) = a_1 + a_2d^2h$ calculated from the data of the first phase sample alone. If $[\mu_x]$ is defined as

$$[\mu_x] = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \text{average number of trees per acre} \\ \text{average sum of tree values } d^2h \text{ per acre} \end{bmatrix} = \begin{bmatrix} 122.7238403 \\ 621801.6947 \end{bmatrix}$$

the first estimator of μ is

$$z_1 = [a]'[\mu_x]$$

where a_1 and a_2 of the vector $[a]' = [a_1 \ a_2]$ are the estimators of the regression coefficients α_1 and α_2 .

The second estimator, to be known here as the second single sampling estimator, is based on the regression function $\hat{y}_3 = r_3(d) = b_1 + b_2d^2$ calculated from the sample data of the first phase alone. If $[b]' = [b_1 \ b_2]$ and

$$[\mu_x] = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \text{average number of trees per acre} \\ \text{average sum of tree values } d^2 \text{ per acre} \end{bmatrix} = \begin{bmatrix} 122.7238403 \\ 11408.69914 \end{bmatrix}$$

the second estimator of μ is

$$z_2 = [b]'[\mu_x]$$

Finally, the third estimator is the double sampling estimator based on the regression function $\hat{y} = r(d) = b_1 + b_2d^2 + b_3d^3 + b_4d^4$ calculated from the combined data of the samples from the two phases, by the double sampling procedure

of Cunia (1982) and Cunia and Michelakackis (1983c). Then, if

$$[\mu_x] = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} = \begin{bmatrix} \text{trees/acre} \\ \text{sum of } d^2/\text{acre} \\ \text{sum of } d^3/\text{acre} \\ \text{sum of } d^4/\text{acre} \end{bmatrix} = \begin{bmatrix} 122.7238403 \\ 11408.69914 \\ 142336.7020 \\ 2148812.849 \end{bmatrix}$$

the third estimator of μ is

$$z_3 = [b]'[\mu_x]$$

where now $[b]' = [b_1 \ b_2 \ b_3 \ b_4]$

The main objective of the present study is to investigate the probability behavior of the double sampling estimator z_3 . But the bias and precision of the first single sampling estimator z_1 , as well as the bias and precision of the height on diameter regression function $\hat{h} = r_2(d)$, may go a long way in explaining the bias, if any and the precision of the double sampling estimator z_3 based on them. Furthermore, when the sample of the second phase is not sufficiently large with respect to the size of the first phase sample, the second single sampling estimator z_2 may be more precise than the double sampling estimator z_3 ; in which case, one may be better off ignoring the information from the second phase sample and use z_2 instead of z_3 .

Because we shall always use, in our present study, regression functions of the same form, there are only twelve estimation procedures; the combinations of four estimation approaches by three estimators. Thus, for each basic sampling method and sample size of the first phase, combined with a basic sampling method and sample size of the second phase (defining a two-phase, two-stage sampling method) there are (i) four sets of up to 100 estimates of regression functions $\hat{y} = r_1(d, h)$, $\hat{h} = r_2(d)$, $\hat{y} = r(d)$ and $\hat{y} = r_3(d)$, one for each estimation approach and (ii) twelve sets of up to 100 estimates z of $\mu = 115.549284$ thousands of pounds of biomass per acre, estimates $(z-\mu)$ of the bias of z and estimates V of the variance of z , one set for each estimation procedure. Because, for some basic sampling methods the number of clusters or the number of trees per cluster may be too small to allow the calculation of regression functions or estimators by some estimation procedure, we cannot be sure that we have exactly 100 estimates z , $(z-\mu)$ and V for each set of the twelve estimators above.

For each estimation procedure and each set of up to 100 estimates z and V , we have calculated the basic summary statistics

$$\bar{z} = \sum z/k, \quad (\bar{z} - \mu) = \sum (z - \mu)/k$$

$$\bar{V} = \sum V/k \text{ and } S_{zz} = \sum (z - \bar{z})^2 / (k-1)$$

where \sum is taken over the $k \leq 100$ elements of each set of the given estimation procedure for the given combination of the basic sampling methods (and sample sizes) of the first and second

phase. We have also calculated additional summary statistics such as, for example, $\sqrt{V/S_{zz}}$, $(S_{zz})/Z$, $t = (Z-\mu)/\sqrt{S_{zz}/k}$ and the number of times μ fell within the 95 and 99 percent confidence intervals $(z \pm t\sqrt{V})$ where $t = 2$ for the 95 and $t = 2.6$ for the 99 percent confidence level.

Analysis Procedure

For a given two-phase, two-stage sampling method (and given number of sample clusters and sample trees per cluster) the 100 simulation runs can be viewed as 100 random experiments (in the statistical sense), where each experiment gives rise to 12 sets of random variables, the estimators z , V , $(z \pm t\sqrt{V})$ etc., one set for each estimation procedure. Under the assumptions of the simulation process the runs are statistically independent and performed under identical conditions. As long as these assumptions are satisfied, and there seems to be no reason to doubt that these assumptions are sufficiently well satisfied, the set of random variables are also statistically independent and identically distributed. To simplify our discussion we shall assume that each individual simulation run generates samples for which the 12 estimation procedures can all be applied. The fact that we may have less than 100 times 12 complete sets of random variables can be explained by the fact that small samples can sometimes be obtained such that estimates cannot be calculated by some estimation procedures.

The sample mean \bar{z} and sample variance S_{zz} calculated from the 100 random variables z are unbiased estimators of the true mean μ_z and variance σ_{zz} of z . Because z is used as an estimator of the mean biomass per acre μ , the statistic $(z-\mu)$ is an unbiased estimator of the bias $(\mu_z - \mu)$ of z as an estimator of μ . When the assumptions of the estimation procedure are satisfied by our population and sample of trees, the bias is expected to be equal to zero. The null hypothesis that the bias is indeed equal to zero can be tested by the statistic $t = (\bar{z} - \mu)/\sqrt{S_{zz}/100}$; this statistic has the approximate t-distribution with 99 degrees of freedom.

Using the tree data of one sample alone, one can also estimate the variance of z by the statistic V . In real life, where data from only one sample is available, V is the only estimator of the variance of z we have. If the assumptions of the estimation procedure are sufficiently well satisfied, then V is a good estimator of σ_{zz} . Consequently, a comparison between V , or better, the average \bar{V} of the 100 estimators V , would furnish information about how well the model assumptions are satisfied for the given estimation procedure (and given population, sampling method and sample size). This comparison can be based on (i) differences between estimators V , or their average \bar{V} , and S_{zz} , (ii) ratio V/S_{zz} or ratio of the averages \bar{V}/S_{zz} , or (iii) differences or ratios of the corresponding standard deviations. We have preferred working with the ratios \bar{V}/S_{zz} .

We have analyzed (i) the bias of z , (ii) the precision of z as measured by the unbiased estimator S_{zz} and (iii) the statistic V as estimator of the variance of z . In addition, we have analyzed the probability behavior of the 95 and 99 percent confidence intervals $(z \pm t\sqrt{V})$ of μ calculated under the assumptions of the estimation procedure by counting the number of times these intervals included μ or happened to fall below or above μ . Because there is an enormous number of combinations of estimation procedures and pairs of samples of the first and second phase, we had to analyze the simulation results by a systematic approach consisting of several steps, the type of analysis performed in one step being conditioned by the results of the analysis of the previous steps.

We shall start with the conclusions reached by Michelakackis and Cunia (1985) when they analyzed the results from one sampling method. They have found that, among other things, (i) the bias, if any, of the double sampling estimator z is negligibly small, whenever the regression equation is suitably selected, (ii) the ordinary least and weighted least squares (OLS and OWLS) approaches lead to estimators that are somewhat better than those obtained by the modified least and weighted least squares methods (MLS and MWLS), (iii) the error of the biomass estimators is grossly underestimated by OLS and OWLS and slightly overestimated by the MLS and MWLS and (iv) the second single sampling estimator z_2 (based on the regression function $\hat{y} = r_2(d)$ calculated from the data of the first phase sample only) is not as good as the double sampling estimator z_3 (based on the regression function $\hat{y} = r(d)$ calculated from the data of both phases). This last conclusion differs somewhat from that reached earlier by Cunia and Michelakackis (1983) when, for their sample data at least, it seemed better to ignore the information from the second phase sample. It is the objective of the present study to further investigate the generality of these conclusions, as additional sampling methods are being considered.

The analysis procedure as used here consists of several main steps. We shall start with the analysis of the effect of (i) number of sample clusters, (ii) number of sample trees per cluster and (iii) estimation within the first basic sampling method applied to each of the first and second phase. With this method, a fixed percentage of trees is selected (from the sample clusters) without replacement and with equal probability. The same type of analysis will be then performed on the data selected by each of the other six basic sampling methods where the fixed percentage is replaced by a fixed number of sample trees per cluster, selected with or without replacement, with equal probability or probability proportional to tree height h , diameter d , basal area d^2 or approximate volume d^2h . Recall that this refers to the data of the first phase only, since the sample of the second phase is still selected by the first basic sampling method.

In the final step we shall compare the conclusions reached in the previous steps by specifically analyzing the differences between the seven basic sampling methods with respect to the bias of z , the precision of z and the estimation of this precision by the statistic V . We shall, thus, draw the overall conclusions about the specific differences between the sampling methods that select, with or without replacement, a fixed number or a fixed percentage of trees per cluster, and with equal or unequal probability.

We were not able to perform a giant-size type of analysis of variance (or covariance) on the entire set of simulated sample data; the processing of statistics derived from millions of samples selected by various sampling methods, various sample sizes and various estimation procedures seems prohibitively complex. Instead, we have preferred using an intuitive, or largely subjective approach based on an ocular analysis of a large number of two and three dimensional tables and graphs, constructed from a representative part of the simulated data. On occasion some simple t-tests were made to support some of the questionable intuitive conclusions. The next two sections will illustrate the application of this analysis procedure.

Analysis of Results - Effect of Sample Size and Estimation Procedure within Basic Sampling Methods

Because of the large number (over 10,000) of all possible combinations of estimation procedures, sample sizes (number of clusters and number of trees per cluster) of the first phase and sample sizes of the second phase, we started with the analysis of a representative set; the combinations of all estimation procedures by all sample sizes of the first phase with the following six sample sizes of the second phase: 11.72 percent of the trees selected from 50 sample clusters, 2.93, 5.86 and 11.72 percent of the trees selected from 200 sample clusters and, finally, 5.86 and 11.72 percent of the trees selected from 400 clusters. It was decided to consider additional combinations only if questions will arise about the conclusions reached from the analysis of the results from the subset above. Fortunately enough, this was seldom necessary.

The statistical results of this analysis were summarized in hundreds of pages of computer output, and further summarized in a large number of tables and graphs. The tables give (i) the number of clusters, the percentage of trees per cluster and the average size n of the sample of the first phase and (ii) the estimates $(\bar{z}-\mu)$ of the bias of z , \sqrt{V} and $\sqrt{S_{zz}}$ of the standard deviation of z , the ratios $\sqrt{V}/\sqrt{S_{zz}}$ of these standard deviations and the sample t-values (to test the null hypothesis that the bias of z is equal to zero). A table contains the statistics of all of the first phase sample sizes for a given sample size of the second phase sample and an estimation procedure. There is a total of 32 tables, the

product of 4 least squares estimation approaches and 8 estimators (two single sampling and six double sampling estimators, one for each sample size of the second phase as stated above). As an example, Table 1 shows parts of two such tables, the statistics of the OLS and MWLS approaches applied to the sample size of the second phase consisting of 11.72 percent of the trees selected from 400 sample clusters. To answer specific questions, the information from the 32 tables was used by the computer to generate several sets of 32 graphs, one graph of each set for each table above.

The first set of graphs shows the average estimates $(\bar{z}-\mu)$ of the bias of z plotted against the average size n , where n denotes the average of the total number of trees per first phase sample, as calculated from the set of up to 100 samples generated by the 100 simulation runs. To facilitate the analysis, the computer drew a smooth curve joining all values $(\bar{z}-\mu)$ of the first phase containing equal number of clusters. Figure 1(a) shows, as an example, the graph corresponding to the OLS statistics of Table 1. An ocular analysis of Table 1 and Figure 1(a) shows that (i) the absolute size of the bias seems to stabilize around the value zero, (ii) in relative terms, say $(\bar{z}-\mu)/\bar{z}$, the bias is small and not significantly different from zero, (iii) the sample estimate of the bias seems to decrease with the increase of the average sample size n and (iv) there seems to be no effect of the number of clusters in the first phase sample.

Approximately the same conclusions can be drawn from the analysis of the remaining 31 graphs. Because of the nested fashion by which the samples were generated, it seems reasonable to draw the overall conclusions by looking at the union of all sample clusters and trees represented by the two samples with largest value n , those of 100 percent of the trees from 15 sample clusters ($n \approx 501$) and 30 percent of the trees from 50 sample clusters ($n \approx 507$). Drawing conclusions from the average bias of all first phase samples of all sizes, gives too much weight to the sample trees of the smaller size samples. For example, the first 5 percent of the trees from the first cluster (of a given simulation run) is contained in the corresponding samples of all other sample sizes.

The main conclusions from the analysis of the entire set of tables and graphs can be summarized as follows. The sample bias is negligibly small, usually less than .5 percent of the value of the estimate \bar{z} . With very few exceptions, the bias is not significantly different than zero; even when it is large and significant, an increase of the sample size n of the first phase would generally make the bias small and not significantly different than zero. The average size of the sample bias does not seem to be affected by (i) the number of clusters or the number of trees per cluster (as long as the overall sample size n remains the same) or (ii) the estimation procedure. It only seems to be affected by the average sample size n of the

Table 1 - Estimates of the bias ($\bar{z}-\mu$), standard deviations ($\sqrt{S_{zz}}$ and \sqrt{V}) and ratio \sqrt{V}/S_{zz} of the double sampling estimators z by the OLS and MWLS estimation procedures. The average sample size n , number of clusters m and percent of trees per cluster p of phase 1 are as shown in the table, while the sample size of phase 2 is that of 11.72 percent of trees selected from each of the 400 sample clusters

First Phase Sample			OLS Estimation Procedure				MWLS Estimation Procedure			
m	p	n	$\bar{z}-\mu$	$\sqrt{S_{zz}}$	\sqrt{V}	\sqrt{V}/S_{zz}	$\bar{z}-\mu$	$\sqrt{S_{zz}}$	\sqrt{V}	\sqrt{V}/S_{zz}
10	5	17	1.69	10.17	7.53	.74	1.61	9.32	8.42	.90
	10	33	1.01	8.10	5.43	.67	2.37	8.18	7.64	.93
	15	50	.69	6.92	4.53	.65	1.21	7.07	6.96	.99
	30	101	.23	6.03	3.33	.55	1.01	5.99	6.55	1.09
	40	134	.32	5.81	2.96	.51	.96	5.74	6.23	1.09
	60	202	.18	5.73	2.49	.43	.45	5.94	6.11	1.03
	100	336	.08	5.31	2.04	.38	.50	5.44	6.01	1.11
15	5	25	.96	7.70	6.11	.79	.42	8.10	8.20	1.01
	10	50	.24	6.16	4.53	.73	.71	6.75	6.49	.96
	15	75	.52	5.52	3.80	.69	1.07	5.81	5.84	1.01
	30	150	.01	5.08	2.81	.55	.51	5.27	5.31	1.01
	40	200	-.02	4.85	2.51	.52	.43	5.01	5.11	1.02
	60	301	-.08	4.73	2.13	.45	.33	5.05	4.92	.97
	100	501	-.09	4.56	1.74	.38	.34	4.92	4.71	.96
20	5	34	.18	6.21	5.44	.87	.21	6.43	6.80	1.06
	10	68	.09	5.14	3.82	.74	.50	5.46	5.62	1.03
	15	101	.32	4.60	3.23	.70	.75	5.16	5.05	.98
	30	202	.29	4.18	2.43	.58	.65	4.46	4.60	1.03
	40	270	.23	4.27	2.17	.51	.65	4.49	4.42	.99
	60	405	.10	4.08	1.85	.45	.63	4.39	4.28	.98
30	5	51	.13	4.60	4.50	.98	.00	4.91	5.51	1.12
	10	101	-.16	3.59	3.25	.91	.18	3.94	4.61	1.17
	15	152	.03	3.36	2.76	.82	.41	3.59	4.25	1.18
	30	304	.11	3.06	2.07	.68	.44	3.48	3.85	1.11
	40	406	.13	3.04	1.86	.61	.46	3.52	3.76	1.07
50	5	85	.34	3.87	3.56	.92	.30	4.20	4.31	1.03
	10	169	.03	3.19	2.61	.82	.42	3.42	3.65	1.07
	15	254	.28	2.92	2.21	.76	.69	3.27	3.40	1.04
	30	507	.11	2.45	1.69	.69	.53	2.85	3.14	1.10

first phase sample; but this is probably due to sampling error.

The shape of the relationship curve of the bias with the sample size n (for a given number of clusters of the first phase) seems to be about the same for all single and double sampling estimators. The height of the curve, however, is not generally the same. This seems to imply that, in some sense, the bias from the component samples of the first and second phase (that generate $y = r_1(d,h)$ and $h = r_2(d)$) is transmitted to the double sampling estimator.

The second set of graphs shows the estimates \sqrt{V} of the standard deviation of z plotted against the average size n of the first phase sample. Recall that (i) V is an estimator of the variance of z calculated from the data of an individual sample, under the basic assumptions of the model defining the estimator z , and (ii) \bar{V} is the average of up to 100 values V of similarly generated samples, by the sampling method and sample size. To facil-

itate the analysis, the points of equal number of sample clusters are joined by a smooth curve. The ocular analysis of the 32 tables and graphs leads to the following main conclusions.

The relationship curve of \sqrt{V} and n has, as expected an inverse-J shape. This is true for all estimation procedures and for all curves joining the points of equal number of clusters of phase 1 sample. Little precision seems to be gained by going above a sample size n of 200; the relationship curve for $n > 200$ becomes almost a horizontal line. The effect of the average size n decreases with the estimate of the standard deviation \sqrt{V} , as expected, but not by much.

For the OLS and OWLS methods that ignore the cluster effect, it seems that one single average curve is sufficiently good to represent the sample with any number of clusters. Figure 1(b) shows this relationship for the OLS data of Table 1. When the estimates are calculated by the MLS and MWLS techniques that take into account the

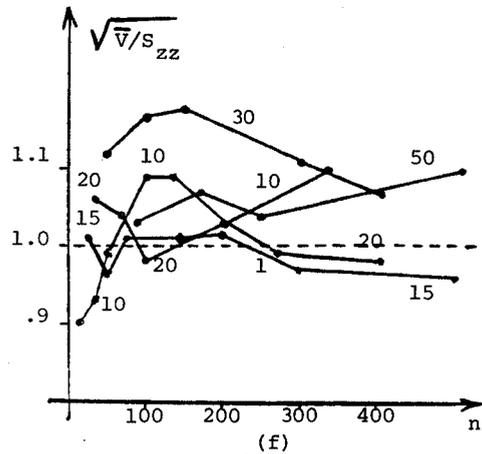
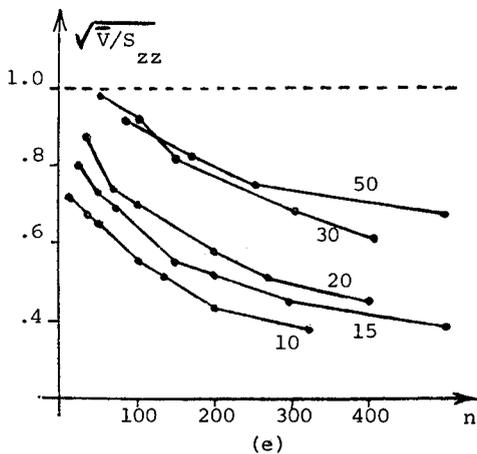
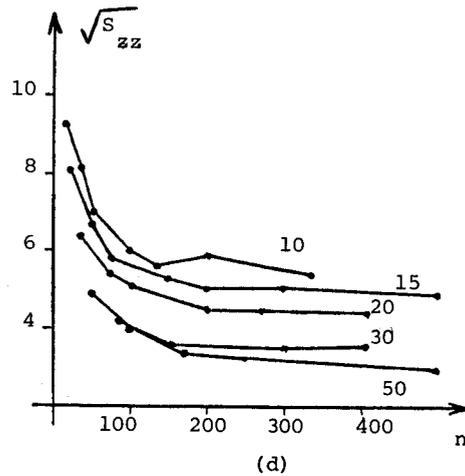
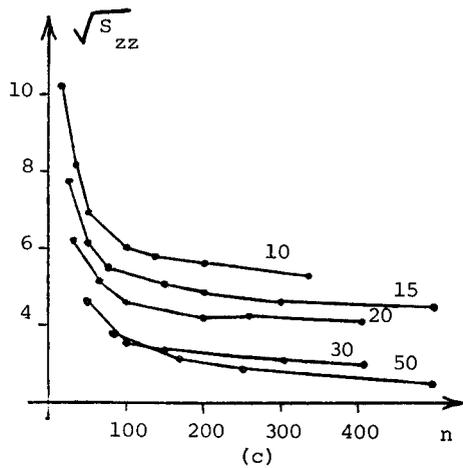
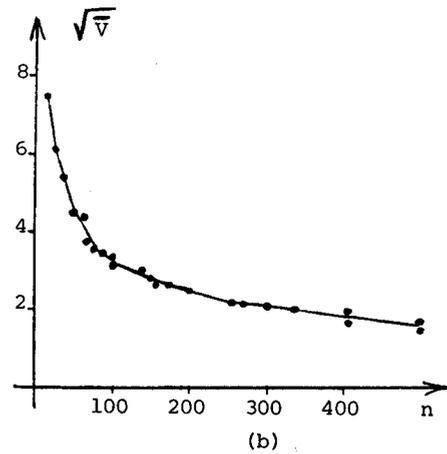
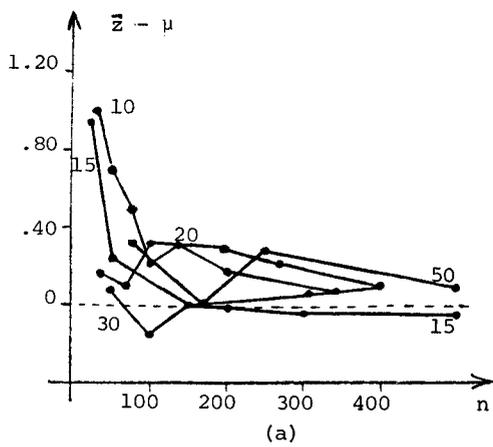


Figure 1 - Some of the relationships of $(\bar{z}-\mu)$, $\sqrt{s_{zz}}$, $\sqrt{\bar{v}}$ and $\sqrt{\bar{v}/s_{zz}}$ with average sample size n of the first phase sample, for the data of Table 1. The number 10, 15, 20, 30 and 50 alongside a curve shows the number m of sample clusters of the first phase sample.

cluster effect, the curves are still the same inverse-J shape but their height depends on the number of clusters contained in the samples of the first phase. As expected, the value of the estimate \sqrt{V} increases with the decrease in the number of clusters; the smaller the number of clusters, the higher the curve. Because the shape of these relationships is the same as that of Figures 1(c) and 1(d), the plot of v against n for the MWLS data of Table 1 is not shown here.

The third set of graphs shows the estimates $\sqrt{S_{zz}}$ of the standard deviation of z plotted against the average size n . With one difference, the conclusions reached about \sqrt{V} holds also true for $\sqrt{S_{zz}}$. Figures 1(c) and 1(d) show, as an example, the relationship curves for $\sqrt{S_{zz}}$ with n , for the OLS and MWLS data of Table 1. The shape of these relationship curves are still of the inverse-J form. But because S_{zz} is an unbiased estimator of the true variance of z , the cluster effect is shown explicitly, as expected for all estimation approaches; the four curves are approximately parallel, with an increase in the standard deviation of z as the number of clusters (for the same overall sample size n of the first phase) decreases.

Because S_{zz} is an unbiased estimator of the true variance of z and there are 12 random variables z (generated by the 12 estimation procedures) for each random experiment (simulation run) we can derive answers to a few additional questions. By comparing first the precision obtained by various estimation procedures we can identify the estimation approach that seems to be the most precise. We can also determine whether there is a minimum sample size for the second phase sample such that, below that size the second single sampling estimator (the one ignoring the second phase sample data) is more precise than the double sampling estimator (which uses the data from both phases). Finally, by comparing the two estimators \bar{V} and S_{zz} , we can determine to what extent V , the only estimator available in real life, is a good and valid estimator of the variance of z .

To identify the estimation procedure that, for a given sample size seems to yield consistently better results, we have constructed additional tables. Because there are six double sampling estimators, and the sample data of the first five are included in the sample data of the sixth, we show here Table 2 listing the standard deviations $\sqrt{S_{zz}}$ by (i) first phase sample size (number of clusters and percent of trees per cluster) (ii) estimation approach and (iii) the second single and the double sampling estimators. In all cases the second phase sample size consists of the selection of 11.72 percent of trees from 400 sample clusters.

An ocular analysis of Table 2 shows that the ordinary weighted least squares (OWLS) estimator seems to yield, most of the time, the estimators of smallest variance, followed closely by the ordinary least squares (OLS) estimators. The two modified least squares (MLS) and weighted least squares (MWLS) estimators are about equally good

but they are both less precise than the OLS estimators. This is not surprising since the range of tree diameters is wider for the OLS (that are based on individual tree values) than the MLS (that are based on the tree averages within the sample clusters).

A brief look at the set of 32 basic tables showed that the double sampling estimators are not always better than the second single sampling estimators (based on the regression function $\hat{y} = r_3(d)$). To better see the combinations of the first and second phase sample sizes for which the single and the double sampling estimators are about equally precise, four more additional tables were constructed, one for each least squares estimation approach. As an example, Table 3 gives the estimated standard deviations of the three types of estimators calculated by the modified weighted least squares approach; the two single sampling estimators and the six double sampling estimators (generated by the six sample sizes of the second phase we have worked with here).

An ocular analysis of the statistics of Table 3 leads to the following conclusions. As expected, the first single sampling estimator based on the regression function $\hat{y} = r_1(d, h)$ is always the most precise. However, when the second phase sample size is sufficiently large, the first single sampling estimator is not much better than the double sampling estimator. For example, if we consider the second phase sample of 5.86 or 11.72 percent of trees from 400 clusters and compare the first single with the double sampling estimators, as shown for MWLS in Table 3, we find that their standard deviations are unusually close; most of the time they differ by about 10 percent. Recall that the population contains 667 non-empty clusters (out of which 400 are in the sample) and the variation between the heights of the trees within a given cluster is relatively small.

Increasing the size of the second phase sample by increasing the percentage of trees subsampled per cluster does not always lead to more precise estimators. For example, if we consider the statistics of Table 3, it appears that better precision is obtained with 5.86 than with 11.72 percent of the trees from both 200 and 400 sample clusters. This may be due to sampling error; the first 5.86 percent of the trees happened to be much less variable than the second 5.86 percent of the trees. But because this is such an unreasonable result, it may also be due to some undetected computational error. On the other hand, an increase in the number of sample clusters (for the same percentage of sample trees per cluster) is always followed by an increase in the precision of z .

Let us now compare the second single sampling with the double sampling estimator. Starting with the second phase sampling method consisting of the selection of 11.72 percent of the trees from 50 sample clusters (a method resulting in an average sample of about 200 trees) we see from Table 3 that, with the exception of the

Table 2 - The standard deviations $\sqrt{S_{zz}}$ of the second single and double sampling estimators calculated by the four least squares estimation approaches. The characteristics of the first phase sample (average sample size n , number of clusters m and percent of trees per cluster p) are as shown and the second phase sample size is that of 11.72 percent of the trees from 400 clusters.

First Phase Sample			Second Single Sampling Estimator				Double Sampling Estimator			
m	p	n	OLS	OWLS	MLS	MWLS	OLS	OWLS	MLS	MWLS
10	5	17	12.15	11.47	11.59	10.55	10.17	8.94	8.77	9.31
10	10	33	10.05	9.50	8.70	9.15	8.10	7.74	7.64	8.18
10	15	50	8.47	7.81	8.29	7.43	6.92	6.87	6.73	7.07
10	30	101	6.92	6.76	6.62	6.50	6.03	5.64	5.85	5.99
10	30	134	6.49	6.61	6.59	6.38	5.81	5.58	5.74	5.74
10	60	202	6.53	6.70	7.00	7.23	5.73	5.62	5.98	5.94
10	100	336	6.07	6.30	6.53	6.79	5.31	5.20	5.56	5.44
15	5	25	10.59	9.16	11.27	10.19	7.70	7.20	8.11	8.10
15	10	50	8.49	7.61	9.10	8.73	6.16	6.23	6.67	6.75
15	15	75	6.96	6.34	7.07	7.07	5.52	5.55	5.50	5.81
15	30	150	5.84	5.69	6.18	6.19	5.08	4.87	5.29	5.27
15	40	200	5.56	5.44	5.98	5.97	4.85	4.67	5.09	5.01
15	60	301	5.29	5.36	5.86	5.87	4.73	4.63	5.10	5.05
15	100	501	5.14	5.22	5.43	5.73	4.56	4.50	4.74	4.92
20	5	34	8.44	7.75	8.73	8.16	6.21	5.92	6.69	6.43
20	10	68	6.83	6.30	7.08	6.86	5.14	5.05	5.53	5.46
20	15	101	5.71	5.27	5.78	5.74	4.60	4.56	4.83	5.15
20	30	202	4.79	4.76	4.84	5.20	4.18	4.05	4.30	4.46
20	40	270	4.80	4.76	4.98	5.23	4.27	4.10	4.42	4.49
20	60	405	4.54	4.48	4.71	4.99	4.08	3.94	4.21	4.39
30	5	51	6.23	5.97	6.99	6.17	4.60	4.54	5.10	4.91
30	10	101	5.00	4.69	4.98	5.06	3.59	3.56	3.82	3.94
30	15	152	4.50	4.07	4.56	4.41	3.36	3.19	3.65	3.59
30	30	304	3.80	3.84	3.82	3.97	3.06	3.16	3.32	3.48
30	40	406	3.78	3.78	3.82	4.06	3.04	3.18	3.31	3.52
50	5	85	5.07	4.70	5.23	4.87	3.87	3.93	4.33	4.20
50	10	169	4.25	3.92	4.30	4.03	3.19	3.01	3.46	3.42
50	15	254	3.48	3.31	3.56	3.38	2.92	2.77	3.17	3.27
50	30	507	3.16	3.17	3.14	3.17	2.45	2.45	2.72	2.85

smallest size samples of the first phase, the single sampling are more precise than the double sampling estimators. But if we consider the remaining second phase sample sizes (consisting of the selection of 2.93 percent or more of the trees from 200 clusters or more), the conclusions change; in all cases, the double sampling estimators seem to be much more precise. This seems to imply that, provided the second phase sample is sufficiently large, the double sampling estimators are better, as they should be. How large the sample should be is difficult to say; it will depend on the population of interest. For our population we have considered additional sizes for the samples of phase 2. For example, when the number of clusters m of the phase 1 sample is higher than 30, the number of clusters of the phase 2 sample should be at least 200, for the double sampling estimator to be more precise; with 150 clusters it was better to ignore the phase 2 sample data and work only with the data from phase 1.

To compare the two estimates \sqrt{V} and $\sqrt{S_{zz}}$ of the standard deviation of z , we have constructed a fourth set of 32 graphs in which the ratios \sqrt{V}/S_{zz} (listed in the 32 basic tables) were plotted against the average sample size n of the first phase sample. The points of equal number of sample clusters were joined together by a smooth curve. As an example, Figures 1(e) and 1(f) show the graphs constructed from the OLS and MWLS data of Table 1. An ocular analysis of the 32 graphs leads to the following main conclusions.

The error of the estimates z of μ derived by the ordinary least and weighted least squares techniques is grossly underestimated by V . This underestimation increases with (i) the decrease in the number of sample clusters and (ii) the increase of the number of sample trees per clusters, given of course, that the average sample size n remains the same. This is to be expected since OLS and OWLS ignore the effect of the

Table 3 - The standard deviations $\sqrt{S_{ZZ}}$ of the eight estimators z calculated by the MWLS approach; m = number of clusters, p = percentage of trees/cluster and n = average sample size.

Single Sampling Estimators			Double Sampling Estimators							
First Phase Sample			Second Phase Sample							
m	p	n	First Estimator	Second Estimator	Number of Clusters and Percent p					
					50	200	200	200	400	400
					11.72	2.93	5.86	11.72	5.86	11.72
10	5	17	8.8	10.5	10.5	9.1	9.4	9.8	8.9	9.3
10	10	34	7.9	9.1	9.7	8.4	8.2	9.0	8.0	8.2
10	15	51	6.5	7.4	9.1	7.4	6.9	7.5	6.8	7.1
10	30	102	5.7	6.5	8.3	6.0	5.9	6.4	5.7	6.0
10	40	136	5.5	6.4	7.8	5.8	5.5	6.1	5.4	5.7
10	60	205	5.6	7.2	7.8	6.0	5.8	6.3	5.7	5.9
10	100	341	5.0	6.8	7.4	5.6	5.5	5.8	5.2	5.4
15	5	23	7.9	10.2	10.1	7.9	7.8	8.4	7.9	8.1
15	10	48	6.6	8.7	8.5	7.0	6.9	7.2	6.7	6.8
15	15	74	5.6	7.1	7.9	6.2	6.1	6.4	5.7	5.8
15	30	151	5.0	6.2	7.7	5.5	5.6	5.9	5.1	5.3
15	40	203	4.8	6.0	7.6	5.4	5.3	5.6	4.8	5.0
15	60	306	4.7	5.9	7.5	5.4	5.3	5.6	5.0	5.1
15	100	512	4.6	5.7	7.4	5.3	5.3	5.5	4.8	4.9
20	5	31	6.1	8.2	8.7	6.2	6.1	6.5	6.0	6.4
20	10	65	5.3	6.9	8.1	5.9	5.7	6.0	5.3	5.5
20	15	99	4.9	5.7	7.8	5.6	5.4	5.7	5.0	5.2
20	30	202	4.2	5.2	7.4	4.8	4.7	5.0	4.3	4.5
20	40	271	4.2	5.2	7.4	4.9	4.7	5.0	4.2	4.5
20	60	406	4.1	5.0	7.3	4.8	4.6	4.8	4.2	4.4
30	5	46	4.6	6.2	8.0	5.1	5.0	5.5	4.7	4.9
30	10	98	3.7	5.1	7.0	4.5	4.4	4.6	3.8	3.9
30	15	149	3.3	4.4	6.8	4.2	4.1	4.3	3.4	3.6
30	30	303	3.1	4.0	6.7	4.0	4.0	4.1	3.1	3.5
30	40	406	3.2	4.1	6.6	4.1	4.0	4.1	3.2	3.5
50	5	77	3.9	4.9	7.1	4.7	4.5	4.8	4.1	4.2
50	10	163	3.1	4.0	6.6	4.2	4.1	4.1	3.4	3.4
50	15	248	2.8	3.4	6.3	4.0	3.9	3.8	3.2	3.3
50	30	505	2.3	3.2	6.1	3.6	3.6	3.5	2.7	2.9

clustering of trees. The total error of the biomass regressions may be viewed as having two additive components; one component due to the variation "within" clusters, the other component due to the variation "between" clusters. When the ordinary least squares methods are used, the two components are affected in the same way; they are both divided by the total number of trees n in the sample, with little effect, if any, on the number of clusters from which they were selected. The proper way would be to have the "between" clusters component affected by the number m of clusters and only the "within" clusters variation affected by the total number of trees n . This is essentially what the modified least squares methods do; the "between" clusters component is roughly divided by the number of sample clusters and the "within" clusters component is roughly divided by the total number of trees.

As the reader can verify by ocular analysis, the error of the estimates z calculated by the two modified least and weighted least squares techniques is sufficiently well estimated by V . The slight overestimation that may exist, is most

probably due to a large extent to the effect of the finite population correction factor (that should normally decrease the value of V), effect that is being ignored here. To see the amount of possible overestimation, consider first the MLS approach. Then (i) \sqrt{V} is about 8 percent higher than $\sqrt{S_{ZZ}}$ when the sampling method of the first phase is 30 percent of trees from 50 clusters (for an average n value of about 500) and (ii) \sqrt{V} is about 5 percent lower than $\sqrt{S_{ZZ}}$ when the sampling method of the first phase is 100 percent of the trees from 15 clusters (for the same average n value of about 500). This yields an average overestimation of about 2 percent. The corresponding overestimation and underestimation percentages for the MWLS approach (shown in Table 1) are about 10 and 4 respectively, for an overall average overestimation of about 3 percent.

It appears from the above discussion that the conclusions reached by Michelakackis and Cunia (1985) have been verified, at least for the first basic sampling method. Together with a few additional conclusions, the results of the analysis of this section can be summarized as follows.

(1) The bias, if any, is small and not significantly different from zero. It does not seem to be affected by the estimation procedure or the size of the second phase sample. The only detectable effect is that of the total number of trees in the first phase sample. As long as this number remains constant, it does not seem to matter the number of sample clusters, or the percentage of trees selected from the sample cluster.

(2) The precision of the estimator z , as measured by the statistic S_{zz} is affected by (i) the estimation procedure (the OWLS seems to be the best approach followed closely by the OLS and then by the two MLS and MWLS approaches that are about equally good), (ii) total number of trees in the first and/or the second phase sample (the larger the sample size, the higher the precision) and (iii) number of clusters for a given total number of trees in the first phase sample (the larger the number of clusters and, thus, implicitly the smaller the number of trees per cluster, the higher the precision, that is, the lower the value of S_{zz}).

(3) The variance of z is grossly underestimated by the sample based statistic V (the only statistic available in the real world), when the OLS and OWLS approaches are being used. On the other hand, the MLS and MWLS approaches result in sufficiently good estimators V of the variance of z , although possibly with a slight overestimation.

(4) For a given, say minimum size of the second phase sample, the estimator z based on the data of the first phase sample alone, (the second single sampling estimator) and the estimator based on the data of the samples from both phases (the double sampling estimator) are about equally precise. If the size goes below this minimum, the second single sampling estimator, which ignores the information from the second phase sample is better. This minimum sample size has not been determined here; it is a function of the sample size of the first phase and the population of interest.

(5) The first single sampling estimator is, as expected, always more precise than the double sampling estimator. This is not surprising. However, this estimator can never be used in real life forest inventory, unless the forest area is very small; it is too expensive to measure the height (in addition to the diameter) of each sample tree. It is precisely for this reason that the second phase sample is being drawn; to reduce the number of trees measured for height. The analysis has shown that as the size of the second phase sample increases, the precision of the double sampling estimator approaches the precision of the first single sampling estimator.

The type of analysis described above, was made for each of the other six basic sampling methods. With small differences, the conclusions were about the same. These differences can be better discussed, however, by comparing the seven basic sampling methods, the topic of the next section.

Analysis of Results: Effect of Basic Sampling Method

The seven basic sampling methods considered here differ mainly by the probability of selection. As stated before, selecting a fixed number (rather than a fixed percentage) of trees from the sample clusters, results in a probability that increases, on the average, with the increase in the tree size. In terms of tree selection from the entire population (not in terms of a given cluster), only the first basic sampling method contains trees selected with equal probability; for all other methods, the probability increases with some measure of tree size. This increase is negligibly small for the basic sampling methods 2 and 3 (where a fixed number rather than a percentage of trees is selected with equal probability from within sample clusters), it is somewhat higher for methods 4 and 5 (where the probability of selection is proportional to tree height and diameter respectively) and the increase in probability becomes quite large with the last two methods 6 and 7 (where the probability of selection within clusters is proportional to basal area and approximate volume).

The difference between the results of the first three basic sampling methods seems to be small, if any; it can hardly be detected from a statistical point of view by the size of our simulated experiment. This means that the difference between selecting a fixed percentage of trees from each sample cluster (to obtain trees sampled with equal probability) and a fixed number of trees (for convenience of sampling and reduction of sampling costs per tree) is small if any. Furthermore, the trees should be selected without replacement (for greater efficiency) since the selection with replacement does not seem to lead to a better estimation of the precision of the statistic z . On the other hand, the differences in the results obtained by the other sampling methods are sufficiently large and should be explicitly identified and analyzed.

To better see the differences in the results obtained by the various sampling methods, we have constructed Table 4 showing the bias ($\bar{z}-\mu$), the precision $\sqrt{S_{zz}}$, the t-values to test the null hypothesis that the bias is equal to zero and the ratio $\sqrt{V/S_{zz}}$ of the two estimates of the standard deviation of z , for the specific combination of largest phase 1 sample of 50 clusters with 10 trees per cluster and the largest phase 2 sample of 400 clusters with 11.72 percent of trees per cluster. For the case of sampling method 1 (where a fixed percentage rather than a fixed number of trees is selected per cluster) we have used 30 percent, or approximately 10 trees, on the average for each of the 50 sample clusters. Recall that 11.72 percent of the trees of phase 2 represents, on the average approximately 4 trees per cluster. The statistics were listed by least squares estimation approach (OLS, OWLS, MLS and MWLS), by sampling method (1, 2, ..., 7) and by type of estimator (first single sampling, second single sampling and double sampling estimator).

Table 4 - The sample bias $(\bar{z}-\mu)$, standard deviation $\sqrt{S_{zz}}$, $t = (\bar{z}-\mu)/\sqrt{S_{zz}/100}$ and ratio $\sqrt{\bar{V}/S_{zz}}$ for the two-phase samples where (i) the first phase consists of 10 trees (or 30 percent for sampling method 1) selected from each of 50 sample clusters and (ii) the second phase sample consists of 11.72 percent of the trees of each of 400 sample clusters.

Estimation Approach	Sampling Method	First Single Sampling				Second Single Sampling				Double Sample			
		$\bar{z}-\mu$	Estimator $\sqrt{S_{zz}}$	t	$\sqrt{\bar{V}/S_{zz}}$	$\bar{z}-\mu$	Estimator $\sqrt{S_{zz}}$	t	$\sqrt{\bar{V}/S_{zz}}$	$\bar{z}-\mu$	Estimators $\sqrt{S_{zz}}$	t	$\sqrt{\bar{V}/S_{zz}}$
OLS	1	.18	2.17	.8	.69	-.06	3.16	-.2	.64	.11	2.45	.4	.69
	2	.30	2.20	1.4	.76	-.41	3.11	-1.3	.72	.23	2.44	.9	.76
	3	-.16	2.37	-.7	.71	-1.55	3.71	-4.2	.61	-.23	2.58	-.9	.72
	4	-.08	2.39	-.3	.77	1.31	3.44	3.8	.71	-.15	2.63	-.6	.76
	5	.48	2.68	1.8	.82	-1.99	3.64	-5.5	.82	.40	2.88	1.4	.82
	6	1.62	4.14	3.9	.71	-2.98	5.20	-5.7	.76	1.53	4.29	3.6	.71
	7	1.48	5.30	2.8	.60	-.42	5.78	-.7	.74	1.40	5.44	2.6	.60
OWLS	1	-.51	2.27	-2.2	.66	-1.21	3.17	-3.8	.54	-.59	2.45	-2.4	.69
	2	-.34	2.24	-1.5	.69	-1.90	3.10	-6.1	.56	-.42	2.44	-1.7	.71
	3	-.61	2.36	-2.6	.65	-2.73	3.33	-8.2	.51	-.69	2.58	-2.7	.67
	4	-.59	2.26	-2.6	.63	.42	3.28	1.3	.51	-.67	2.51	-2.7	.65
	5	-.29	2.20	-1.3	.61	-2.35	3.05	-7.7	.50	-.37	2.43	-1.5	.64
	6	.00	2.02	.0	.60	-2.37	2.85	-8.3	.51	-.08	2.25	-.4	.65
	7	-.33	2.10	-1.6	.55	.01	2.85	.0	.51	-.41	2.35	-1.7	.60
MLS	1	.20	2.19	.9	1.15	-.01	3.14	-.0	1.09	.42	2.72	1.5	1.08
	2	.33	2.21	1.5	1.22	-.43	3.13	-1.4	1.15	.55	2.75	2.0	1.13
	3	-.10	2.41	-.4	1.19	-1.37	3.49	-3.9	1.13	.11	2.93	.4	1.10
	4	-.12	2.44	-.5	1.29	1.18	3.17	3.7	1.37	.09	2.99	.3	1.17
	5	.79	3.27	2.4	1.28	-1.82	4.64	-3.9	1.26	.99	3.74	2.6	1.18
	6	2.34	5.95	3.9	1.03	-2.46	8.53	-2.9	1.01	2.54	6.26	4.1	1.00
	7	2.15	7.18	3.0	.95	-.56	9.24	-.6	1.03	2.35	7.45	3.2	.93
MWLS	1	.30	2.33	1.3	1.17	-.08	3.17	-.3	1.08	.53	2.85	1.9	1.10
	2	.46	2.39	1.9	1.17	-.66	3.24	-2.0	1.05	.68	2.88	2.4	1.11
	3	.10	2.76	.4	1.08	-1.52	3.59	-4.2	.99	.32	3.22	1.0	1.04
	4	-.16	2.47	-.6	1.12	1.12	3.36	3.3	1.03	.07	2.95	.2	1.07
	5	-.12	2.23	-.5	1.16	-2.74	3.15	-8.7	1.04	.10	2.69	.0	1.12
	6	-.06	2.23	-.3	1.13	-4.22	3.24	-13.0	1.07	.16	2.74	.6	1.08
	7	-.57	2.37	-2.4	1.06	-2.44	3.23	-7.6	1.11	-.35	2.87	-1.2	1.03

Of course, we have analyzed many other combinations of sample sizes of the first and second phase. Because (i) the combination of the two largest samples is the union of most of the other smaller samples (recall that the samples of a given simulation run are drawn in a nested fashion), (ii) the general conclusions derived from the analysis of combinations of samples of smaller size are about the same and (iii) the available space for more tables is rather limited in this paper, we have preferred showing only some of the most important results in a single Table 4. Although for the analysis that follows we shall refer to Table 4, the conclusions we have reached are much more general and refer to samples of all sizes.

Let us start with the bias $(\bar{z}-\mu)$ of the statistic z as an estimator of the parameter μ . An ocular analysis of Table 4 shows that the results are confusing, sometimes inconclusive and it is rather difficult to draw any general con-

clusions. Considering first the double sampling estimators we see that with one exception, the sample bias is small (usually less than .5 percent of the mean μ) and not significantly different from zero. The size of the bias is little, if at all affected by the estimation approach (ordinary or modified, least or weighted least squares). The only exception is with the OLS and MLS estimators of the sampling methods 6 and 7 where the bias is relatively large (about 1.5 to 2.0 percent of the mean μ) and significantly different from zero. Recall that, for methods 6 and 7, the probability of tree selection is proportional to basal area and approximate volume respectively. Because the bias is small for these same methods when OLS and MLS are replaced by the corresponding OWLS and MWLS approaches, it appears that by properly weighing the sample data (less weight for the large trees sampled with higher probability) the size of the bias may be reduced.

Exactly the same conclusions can be drawn from the analysis of the bias of the first single sampling estimator. The correlation between the pairs of values of the bias by the single and double sampling methods is extremely high. This is to be expected, since the same regression function $\hat{y} = r_1(d, h)$ is used to derive (i) the first single sampling estimator and (ii) the regression function $\hat{y} = r(d)$ on which the double sampling estimator is based. As the second phase regression function $\hat{h} = r_2(d)$ is calculated from a large sample of 400 clusters (selected without replacement from 667 non-empty clusters of the population) it is not surprising to see that the error of $\hat{y} = r(d)$ has as its main source the error of $\hat{y} = r_1(d, h)$.

On the other hand, the bias of the second single sampling estimators is significantly different from zero, with the exception maybe of the sampling methods 1 and 2, where the probability of tree selection from within the sample clusters is the same for all trees. This may be due to (i) the form of the regression function we have used ($\hat{y} = r_3(d) = \beta_1 + \beta_2 d^2$) that was good in the preliminary study of Michelakackis and Cunia (1985) may no longer be good when applied to sample data collected by other sampling methods) or (ii) simply to inherent sampling error (we have obtained an unusual sample). The bias does not seem to be affected by the least squares estimation approach (OLS, OWLS, MLS and MWLS, they all seem to yield the same bias) but the probability of tree selection seems to affect the bias (sampling with equal probability seems to yield smaller bias, often not significant, while sampling with probability proportional to basal area seems to yield the bias of largest (absolute) value).

A final note about the bias. Because the seven sets of 100 simulation runs (one set for each basic sampling method) are statistically independent (each set was made with a different random start), we may view the seven sample values of the bias ($\bar{Z} - \mu$), one value for each set, as seven estimates of the bias which is zero under an appropriately defined null hypothesis (that $\mu_Z = \mu$ for all seven sampling methods). Based on the theory of binomial distributions, one may then state that, for a given least squares estimation procedure, the bias is significantly different from zero, whenever exactly six or exactly seven sample values of the bias are of the same sign. This is the case with seven out of twelve estimation procedures of Table 4. This seems to imply that (i) the bias of the second single sampling estimator is negative for all estimation approaches, (ii) the bias of the first single sampling and double sampling estimators by the OWLS approach is also negative and finally (iii) the bias of the remaining OLS, MLS and MWLS approaches for the first single sampling and double sampling estimators is positive. However, this last conclusion, as well as the two preceding conclusions are somewhat questionable and should be accepted only with some reservations.

Consider now the precision of z , expressed as the standard deviation (error) of z estimated by $\sqrt{S_{zz}}$. An ocular analysis of Table 4 leads to the following conclusions. Starting with the ordinary and the modified weighted least squares approaches (OWLS and MWLS) one may conclude that, for both single and double sampling estimators, we have (i) the ordinary weighted least squares estimators (OWLS) are more precise than the corresponding modified estimators (MWLS), (ii) within the same estimation approach, the seven sampling methods yield estimators z of about the same precision and (iii) as expected the double sampling estimators are more precise than the corresponding second single sampling estimators but not as precise as the corresponding first single sampling estimators.

It is worth mentioning here that the double sampling estimator is not always more precise than the corresponding second single sampling estimator. For some combinations, when the size of the second phase sample is not sufficiently large with respect to the first phase sample size, the second single sampling estimator is better. In all these cases it may be worth ignoring the information from the second phase sample and base the estimate of μ on the data from the first phase alone.

The same type of ocular analysis performed on the ordinary and the modified least squares methods (OLS and MLS) leads, with one major exception, to the same basic conclusions. The exception is with the sampling methods for which the probability of tree selection is proportional to basal area (method 6), proportional to approximate volume (method 7) and sometimes proportional to diameter (method 5) where the precision of z is much lower than that of the other sampling methods. This is difficult to explain, and for this reason this result should only be accepted with some reservations. Additional research may be necessary to confirm or to reject this conclusion.

Finally, an overall view of Table 4 seems to suggest that, with the exception of the basic sampling methods 6 and 7 (and possibly 5), the OWLS estimator is the most precise, followed closely by the OLS and then by the MLS and MWLS approaches.

Let us turn finally to the values of the ratio \sqrt{V}/S_{zz} . A simple look at Table 4 shows that the error of z is grossly underestimated for the OLS and OWLS, and slightly overestimated for the MLS and MWLS approaches. On the average and for the sample sizes considered in Table 4, the value of the ratio \sqrt{V}/S_{zz} is about .70 - .75 for OLS, about .55 - .65 for OWLS, about 1.10 for MLS and finally about 1.07 for MWLS.

Concluding Remarks

The objectives of this study were those of investigating the validity of statistical inferences one makes, when biomass regression functions are estimated by the ordinary least squares techniques (or techniques modified to take into account the effect of the sampling method), applied to sample tree data selected by a two-phase, two-stage sampling design. The basic procedure of calculating the regression function $\hat{y} = r(d)$ of tree biomass y on tree diameter d , when one is given estimates of the regression functions $\hat{y} = r_1(d, h)$ of biomass on diameter and height and $\hat{h} = r_2(d)$ of height on diameter calculated from the tree data of the first and second phase respectively, is briefly described in the paper. For more details, the reader is referred elsewhere.

We have considered seven basic sampling methods (all of the same two-phase, two-stage type) that differed only by the second phase procedure of subsampling trees from the first stage sample clusters. Within each sampling method we have varied the number of sample clusters and the number of trees selected from these clusters. We have also considered twelve estimation procedures, the combinations of four estimation approaches (OLS, OWLS, MLS and MWLS) with three types of estimators (based on the regression functions $\hat{y} = r_1(d, h)$ of the first phase, $\hat{y} = r_2(d)$ of the second phase and $\hat{y} = r(d)$ of the combined data of the two phases).

Most of the results we have found here are those we expected. For example, we expected the bias to be small and not significantly different than zero, most of the time; the precision to increase with the overall increase of the sample size (total number of sample trees), or within a given sample size, to increase with the increase of the number of sample clusters (or the decrease of the average number of sample trees per cluster); the statistic V to grossly underestimate the variance of z for OLS and OWLS and be approximately right for MLS and MWLS; the weighted least squares procedures to be more precise than the least squares, and the ordinary least squares techniques to yield better estimates of μ than the modified ones; the first single sampling to be more precise and the second single sampling to be most of the time less precise than the double sampling estimators. One result, however, may need emphasizing here in more explicit terms. The method suggested by Cunia (1982) to estimate a biomass regression function (and its error) by double sampling techniques (in our case two-phase, two-stage designs) is essentially sound, provided that the first and second phase regression functions on which it is based have been properly derived and their error properly estimated.

We have also found the following completely unexpected results.

(1) The least squares method is much more robust than we ever thought. It generated estimates that were only slightly less precise than those obtained by the weighted least squares method, the standard technique generally used to estimate biomass regression functions. If the construction of biomass tables is for the purpose of estimating average biomass per tree or unit area, it does not seem to matter much one way or the other, whether the tables are constructed by least or by the weighted least squares. This conclusion may not necessarily apply, however, when the regression functions are estimated for (i) a general use outside the inventory estimation problem considered here or (ii) the application to a limited range of tree size, as for example, to the calculation of the biomass of the trees above 12 inches of diameter.

(2) It is generally thought that sampling trees with probability proportional to some measure of tree size (basal area or volume) leads to better estimates of the biomass regression functions. It is known that the conditional variance of the tree biomass is approximately proportional to tree basal area (d^2) or volume (d^2h). Then, it seems intuitively right to assume that including in the sample a proportionally higher number of large trees would increase the precision of the regression estimator. Cunia (1979) did not dispute the fact that better precision may be obtained when larger trees are sampled with higher probability. But he presented arguments in the sense that proceeding this way, one may introduce a bias of unknown size.

The present simulation study showed that the introduction of a bias in the estimation of μ is a real possibility, especially true when the least squares procedures (OLS and MLS) are being used. It was surprising to see that, the sample bias, which in Table 4 was about 1.5 - 2.0 percent of the mean μ , was much higher for other, smaller sample sizes. But the real surprise was to see that the precision of the estimates did not improve with the increase of the probability of tree selection with the tree size. On the contrary, there was a drastic reduction in this precision for some sample sizes and estimation procedures. Furthermore, to measure the biomass of a large tree may be much more expensive than to measure the biomass of a small tree. Consequently real arguments can be brought in favor of sampling with equal probability. However, because it seems to go against our intuition, we should be more cautious and accept this conclusion with some reservations. Further research may be necessary and additional evidence must be gathered before reaching a definite conclusion in favor or against sampling with probability proportional to size.

Of course, the conclusions drawn from this study are valid, in the strict statistical sense, only for our tree population. As it was constructed from real world data, this population is, in some sense, representative of real world populations. Consequently, we feel that our conclusions are sufficiently general to go beyond the narrow application to one, somewhat artificially constructed forest tree population.

Acknowledgements

This paper is based on research funded by the Research Foundation of the State University of New York, the United States Department of Agriculture Forest Service and the Department of Energy, Grant No. 23-524.

Literature Cited

- Briggs, E. F.; Cunia, T. Effect of cluster sampling in biomass tables construction: linear regression models. Canadian Journal of Forest Research 12: 255-263; 1982.
- Cunia, T. On tree biomass tables and regressions: some statistical comments. In: 1979 forest resource inventories workshop proceedings, W. E. Frayer, (Ed.) Colorado State University, Fort Collins, CO; Vol. II, 629-642; 1979.
- Cunia, T. Cluster sampling and tree biomass tables construction. In: Interdivisional Proceedings, 17th IUFRO World Congress, September 6-12, 1981, Kyoto, Japan; 1981.
- Cunia, T. On the error of tree volume tables and its effect on the precision of forest inventory estimates. In: Statistics in theory and practice: essays in honor of Bertil Matern. B. Ranney (Ed.). Swedish University of Agricultural Sciences, Section of Biometry, S-90183, Umea, Sweden; 1982.
- Cunia, T. Evaluating errors of tree biomass regressions by simulation. In: Proceedings of the workshop of "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Cunia, T.; Michelakackis, J. A method to construct a forest biomass population model. In: Proceedings, Renewable resource inventories for monitoring changes and trends. J. F. Bell and T. Atterbury (Eds.), Oregon State University, Corvallis, OR; 1983a.
- Cunia, T.; Michelakackis, J. On the error of tree biomass tables constructed by a two-phase sampling design. Canadian Journal of Forest Research. 13: 303-313; 1983b.
- Cunia, T.; Michelakackis, J. A Monte Carlo technique for generating total height of forest trees. Faculty of Forestry Miscellaneous

Publication Number 4 (ESF 84-018), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1984a.

Cunia, T.; Michelakackis, J. Constructing forest biomass populations for simulated sampling. Faculty of Forestry Miscellaneous Publication Number 5 (ESF 84-019), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1984b.

Cunia, T.; Michelakackis, J.; Lee, S. Generating total tree heights by a Monte Carlo technique. In: Proceedings, 1983 Southern forest biomass workshop, June 15-17, 1983, Charleston, SC, R. F. Daniels and P. H. Dunham (Eds.) USDA Forest Service, Southeastern Forest Experiment Station, Asheville, SC; 1984.

Michelakackis, J.; Cunia, T. Construction of biomass tables by double sampling: preliminary results of a simulation study. In: Proceedings, Use of auxiliary information in natural resource inventories, October 1-2, 1985, Blacksburg, VA. R. G. Oderwald, H. E. Burkhardt and T. E. Burk (Eds.), Society of American Foresters Publication No. SAF 86-01; 1985.

USING SIMULATION TO EVALUATE VOLUME EQUATION
 ERROR AND SAMPLING ERROR IN A TWO-PHASE DESIGN^{1/}

David C. Chojnacky

Research Forester, Intermountain Research Station,
 Forest Service, U.S. Department of Agriculture,
 Ogden, UT 84401

Three volume estimators were evaluated using computer simulation on a population of 150 timber plots. The estimators were designed in a two-phase fashion where a large sample of trees were measured for diameter and height and a subsample were measured for volume. Bias and variance performance were compared among estimators.

Introduction

A simple linear volume equation ($V = a + b \cdot D^2H$) is often used in timber surveys without regard to its potential to inject volume prediction errors. In essence, the volume equation is assumed to predict without error. However, error can be a problem and can be dealt with by devising a model-based estimator using the D^2H volume equation. Volume prediction error can then be quantified and added to sampling error in a timber survey.

Bose (1941-42) first designed such an estimator to include the effects of linear regression equation error, with sampling error, in cinchona bark survey for India's forests. Bose used a small sample to estimate parameters for an equation relating cinchona bark volume to an auxiliary variable, and a larger sample to estimate the population mean of the auxiliary variable. The sample variance for his estimator was simplified by assuming the existence of a bivariate normal distribution between cinchona bark volume and the auxiliary variable. Khan and Tripathi (1967) extended Bose's work to use of a multiple regression equation in a two-phase sampling design. They assumed a multivariate normal distribution between the auxiliary variables and the variable of interest to derive variance estimators. Cunia and Michelakackis (1983) applied Khan and Tripathi's method to timber volume sampling but without the multivariate normal assumption for variance derivation. Instead, they used an approximation formula for a general function of random variables to compute a variance for their estimator (Kempthorne and Folks 1971). Pfefferman and Nathan (1977) extended regression-equation sample designs to include selection of sample elements in clusters. But they used a Bayesian approach that required known distributions for the regression parameters associated with the auxiliary variables.

In my Ph.D. dissertation (Chojnacky 1985), I devised five estimators to sample timber volume using a D^2H equation in two-phase cluster design. The auxiliary variable (D^2H) was determined in phase one, and the volume equation parameters were estimated in phase two. The estimators were similar to those described above, except variances were derived without making distribution assumptions about the auxiliary variable or regression parameters (slope and intercept). For this paper, I evaluated three of the most promising estimators (Table 1) and their variances (Table 2) in a computer simulation of 500 repeated samples from a tree volume population. Estimator 1 was model-unbiased and required the fewest assumptions for variance derivation. However, it was a special case of the two-phase design where phase-one and phase-two sample sizes were equal.

Table 1.--Three estimators for computing mean volume per cluster in a two-phase design. (See list of symbols at end of text for complete symbol definitions.)

Estimator	Cluster sample mean
1	$\hat{\bar{Y}}_1 = \frac{1}{n} \sum_{i=1}^n (\hat{\alpha}_1 T_i + \hat{\beta}_1 X_i)$
2	$\hat{\bar{Y}}_2 = \frac{1}{m} \sum_{i=1}^m (\hat{\alpha}_1 T_i + \hat{\beta}_1 X_i) \frac{\frac{1}{n} \sum_{i=1}^n X_i}{\frac{1}{m} \sum_{i=1}^m X_i}$
3	$\hat{\bar{Y}}_3 = \frac{1}{n} \sum_{i=1}^n (\hat{\alpha} T_i + \hat{\beta} X_i)$

where

- T_i = total number of trees in cluster i
- X_i = sum of D^2H values in cluster i
- n = number of clusters in phase-one sample
- m = number of clusters in phase-two sample
- $\hat{\alpha}_i, \hat{\beta}_i$ = model parameters estimated for cluster i
- $\hat{\alpha}, \hat{\beta}$ = model parameters estimated from all cluster data combined.

^{1/}Paper presented at a national workshop on Tree Biomass Regression Functions and Their Contribution to the Error of Forest Inventory Estimates, Syracuse, NY, May 26-30, 1986.

Table 2.--Sample variances for the three estimators partitioned into variation due to phase 1 sampling, phase 2 sampling, and model uncertainty. (See list of symbols at end of text for symbol definitions.)

Estimator	Classical sample variance		
	Partition 1	Partition 2	Partition 3
1	$v(\bar{Y}_1 - \bar{Y}_{..}) = \left(1 - \frac{n}{N}\right) \frac{s^2 Y_1}{n}$		$+ \frac{1}{n} \sum_{i=1}^n T_i \sigma_i^2 \left[\frac{T_i}{nt_i} + \frac{T_i (\bar{X}_i - \bar{X})^2}{n(t_i - 1) s_{x_i}^2} - \frac{1}{N} \right]$
2	$v(\bar{Y}_2 - \bar{Y}_{..}) = \left(1 - \frac{n}{N}\right) \frac{s^2 Y_2}{n}$	$+ \left(1 - \frac{m}{n}\right) \frac{s^2 d}{m}$	$+ \frac{1}{m} \sum_{i=1}^m T_i \sigma_i^2 \left[\frac{T_i}{mt_i} + \frac{T_i (\bar{X}_i - \bar{X})^2}{m(t_i - 1) s_{x_i}^2} - \frac{1}{N} \right]$
3	$v(\bar{Y}_3 - \bar{Y}_{..}) = \left(1 - \frac{n}{N}\right) \frac{s^2 Y_3}{n}$		$+ \frac{1}{n} \sum_{i=1}^n T_i \sigma_i^2 \left[\frac{T_i}{\bar{m}t} + \frac{T_i (\bar{X}_i - \bar{X})^2}{(\bar{m}t - 1) s_x^2} - \frac{1}{N} \right]$

In this application of two-phase sampling, volume equation parameters were estimated from the phase-two sample. Because those parameters require costly field measurements, it was more practical for field applications to use an estimator with a phase-two size much smaller than the phase-one sample. Estimators 2 and 3 were biased but did allow a smaller phase-two sample size.

Objectives in using the repeated samplings were to:

1. Determine each estimator's bias
2. Compute a simulation variance for each estimator
3. Construct confidence intervals for each estimator

Data

Tree volume inventory data from Idaho, Wyoming, and Utah were used to evaluate the three estimators. The data, treated as a population, included 150 clusters of trees sampled for cubic meter volume from the Ashley, Challis, Salmon, and Targhee National Forests (Fig. 1). The volume clusters were a subsample of a large stratified random sample covering the entire forests. From each cluster, at least six trees 12.7 cm diameter at breast height (d.b.h.) and larger were selected proportional to d.b.h. along a transect. These trees were felled and measured for inside bark volume, d.b.h., and total height.

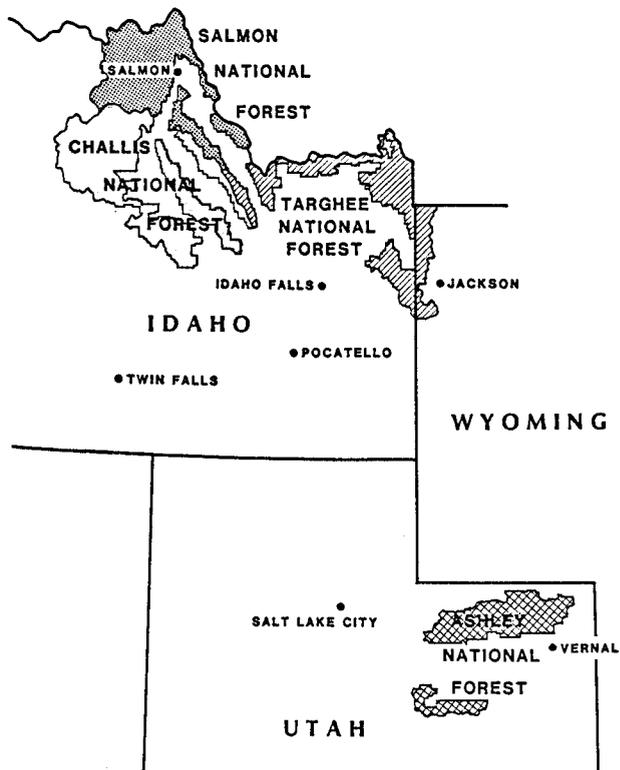


Figure 1.--Area map of the Salmon, Challis, Targhee, and Ashley National Forests.

Eight species--white bark pine, subalpine fir, white fir, aspen, Douglas-fir, Engelmann spruce, lodgepole pine, and ponderosa pine--were sampled. No more than four species occurred in a single cluster. In more than half the clusters, only one species occurred. Figure 2 illustrates the frequency distribution of total volume per cluster for the 150 clusters.

Simulation And Computation

For computer simulation, the population of 150 clusters was sampled 500 times in two phases (Table 3). In phase one (sample n) a sampling fraction of 0.5 was used to obtain 75 clusters from the population. For phase two (sample m) three sampling fractions of 0.13, 0.27, and 0.53 were used to obtain samples of size 10, 20, and 40 clusters from the first-phase sample. Samples in both phases were randomly selected without replacement. Sample sizes selected represented a compromise between choosing the most interesting comparisons and keeping within a limited computer budget. Because volume equation parameters were determined in phase two, this phase was most interesting for sample size comparisons. Therefore, I chose three sample sizes for phase two. I saw little advantage in varying the phase-one sample size, so it was kept constant. Given sample sizes in both phases, 500 repeated samples was the maximum allowed by the computer budget.

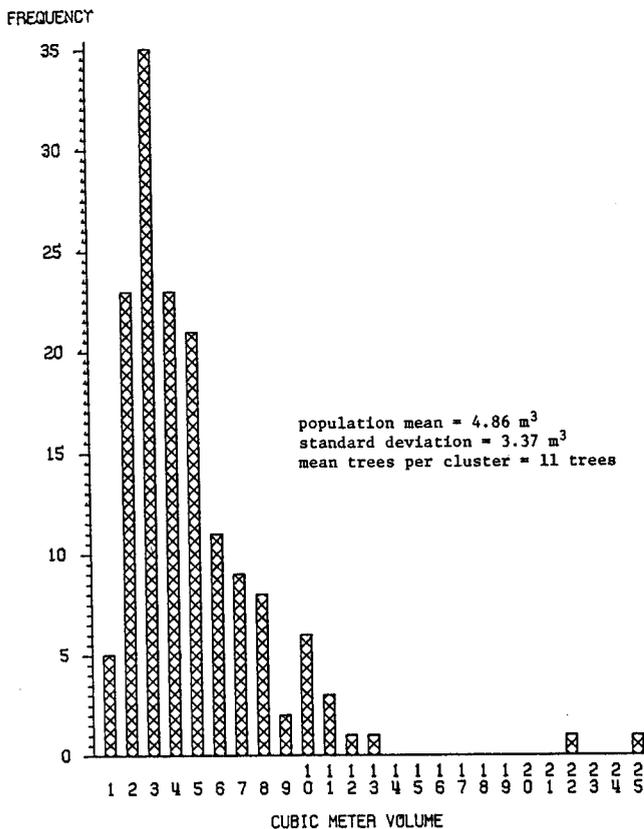
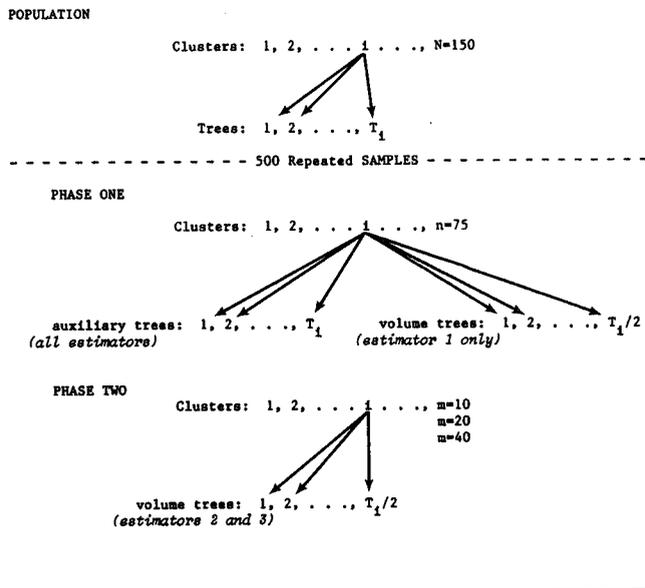


Figure 2.--Frequency distribution of cubic meter volume per cluster for the population.

Table 3.--Diagram of the repeated sampling process.



The auxiliary variable (d.b.h. squared times height) for the volume regression equation was obtained from every tree selected in phase one. Also in phase one, half the trees (t_i) in each cluster were selected to estimate volume equation parameters. Equation parameters for estimators 2 and 3, were estimated from half the trees in each cluster selected in phase two. Each estimator and its sample variance were computed 500 times. In addition, a single simulation variance was computed for each estimator:

$$\text{Simulation variance} = \sum_{j=1}^R (\bar{y}_{ij} - \bar{y}_{i.})^2 / (R-1) \quad (1)$$

where \bar{y}_{ij} = mean volume per cluster for the i^{th} estimator for the j^{th} repeated sample
 $i = 1, 2, 3$ for the 3 estimators
 $R =$ number of repeated samples.

Simulation Results

Frequency distributions of the 500 sample means for each estimator appeared to follow the normal distribution (Fig. 3). I saw little difference among these distributions for the phase-two sample sizes ($m = 10, 20, \text{ or } 40$). Bias for each estimator was determined by comparing the population mean (4.856 m³ per cluster) to each estimator's mean of 500 sample means. Less than 1 percent bias was found for each estimator among all sample size combinations (Table 4).

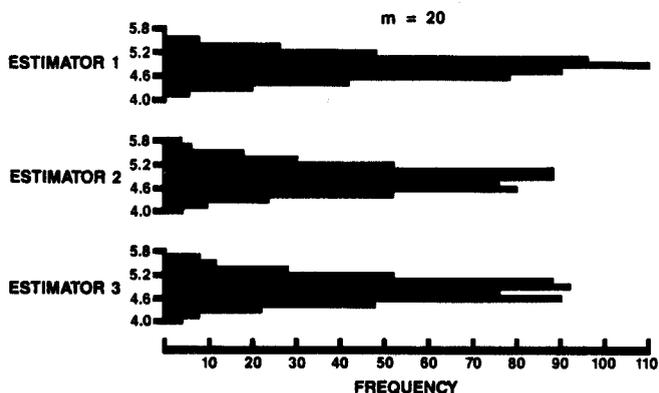


Figure 3.--Frequency distributions of sample means for m=20.

Variances were first examined by comparing each estimator's simulation variance to its respective mean sample variance (Table 5). These means from the sample variance distributions were closest to their respective simulation variance for estimator 1. For estimators 2 and 3, the difference between the two variances decreased as sample size increased. Because most sample variance frequency distributions were highly skewed (Fig. 4), distribution quantiles were compared to the simulation variance. Comparison of quantile plots (in Fig. 5) clearly showed the sample variance of estimator 1 to be the most reliable variance estimator throughout the distribution of repeated samples. The sample variance for estimator 3 was next in reliability in considering total performance for all sample sizes of m.

Table 4.--Bias of the three estimators.

Estimator	Sample sizes		Bias ^{a/}
	n	m	
			Percent
1	75	-	0.2
		-	-.1
		-	.2
2	75	10	.6
		20	.0
		40	.3
3	75	10	.8
		20	-.1
		40	.1

^{a/} Percent bias is defined as the mean of the estimator (sample mean) distribution minus the population mean, divided by the population mean.

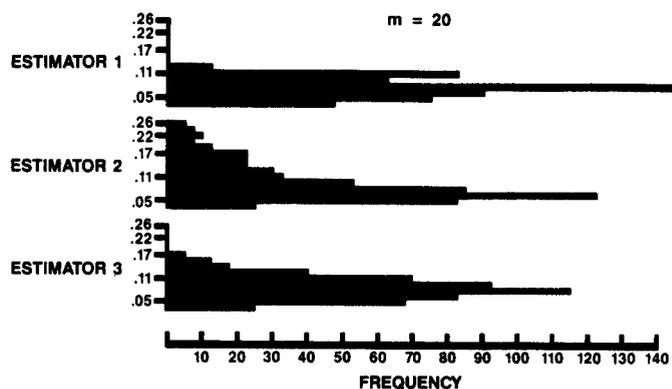


Figure 4.--Frequency distributions of estimator sample variances for m=20. (Note the nonuniform vertical scale.)

Confidence intervals were constructed using both the simulation and sample variances. For the simulation variance, confidence intervals were based on the normal distribution without the need for a degrees of freedom (df) variable. This was because the simulation variance was computed from a large distribution of sample means that approximate the normal distribution (see Fig. 3). On the other hand, the sample variance involved small sample sizes for m, requiring use of the t-distribution and a df variable. Selecting appropriate df for the t-distribution confidence intervals did present a slight problem. While a standard t-table is designed for one df variable, estimators in this study presented up to three. In Table 2, the variances of the estimators are partitioned into

Table 5.--Simulation variance and mean sample variance for the three estimators.

Estimator	Sample sizes		Simulation variance	Mean of sample variance distribution
	n	m		
1	75	10	0.082	0.079
		20	.072	.077
		40	.073	.078
2	75	10	.134	.117
		20	.102	.092
		40	.080	.083
3	75	10	.129	.099
		20	.093	.086
		40	.080	.082

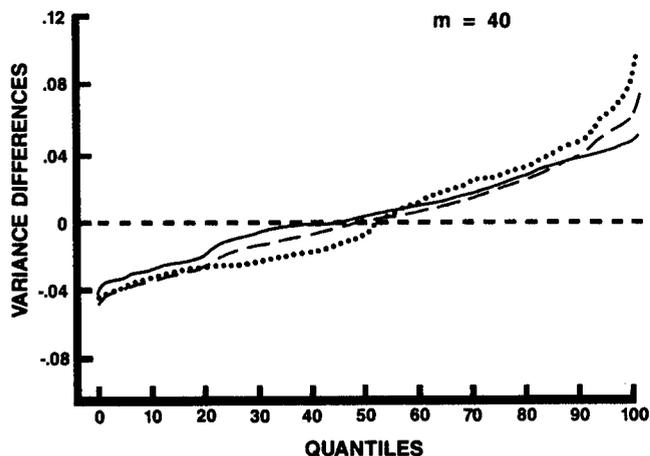
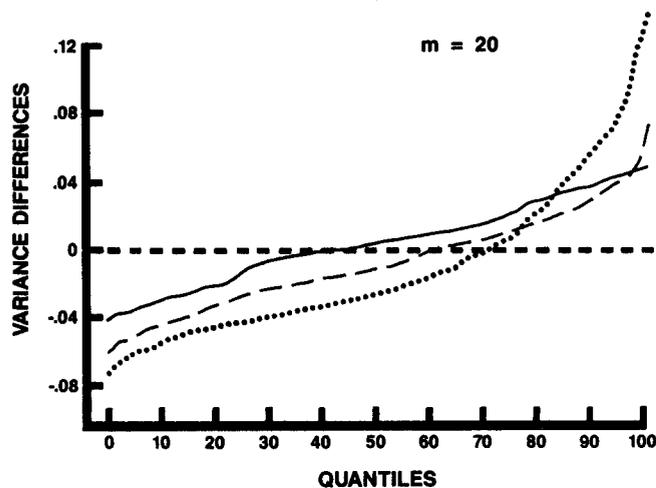
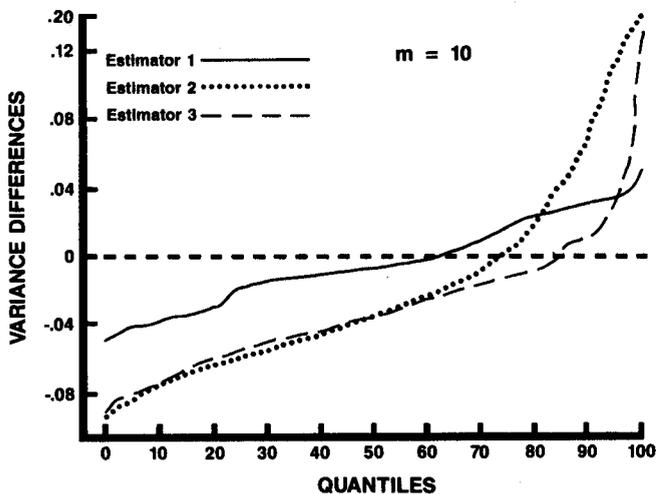


Figure 5.--Difference between the sample variance of the estimator and its corresponding simulation variance for quantiles of the sample variance distribution for m equal to 10, 20, and 40. (Note the nonuniform vertical scale for $m=10$.)

three terms representing three sources of variation and three df values. The following formula (Cochran 1977, p. 96) was used to compute the combined degrees of freedom:

$$df_{ij} = \frac{(v_{1ij} + v_{2ij} + v_{3ij})^2}{\frac{v_{1ij}^2}{df_{1ij}} + \frac{v_{2ij}^2}{df_{2ij}} + \frac{v_{3ij}^2}{df_{3ij}}} \quad (2)$$

where

- i = estimators 1 through 3
- j = samples 1 through 500
- df = degrees of freedom
- v_{1ij} = variance due to phase-one sampling (1st partition, Table 2)
- v_{2ij} = variance due to phase-two sampling (2d partition, Table 2)
- v_{3ij} = variance due to the model (3d partition, Table 2)
- $df_{1ij} = \begin{cases} n-1 & \text{for } i=1 \\ m-1 & \text{for } i=2,3 \end{cases}$
- $df_{2ij} = m-1$ for all i
- $df_{3ij} = \begin{cases} n-1 & \text{for } i=1 \\ m-1 & \text{for } i=2,3. \end{cases}$

To computerize the confidence interval computations, t -values were computed from an asymptotic expansion of the inverse of the cumulative student's t -distribution given by Abramowitz and Stegun (1964, p. 949). Only four terms of the asymptotic expansion had to be used because resulting t -values corresponded almost exactly (to three significant digits) to those given in a standard t -table for degrees of freedom larger than three.

Confidence intervals of 95 percent were computed for each estimator using both the simulation and sample variances. The number of confidence intervals containing the population mean was tallied (Table 6). For the simulation variance, the percentage of confidence intervals containing the population mean was close to the theoretical 95 percent. On the other hand, confidence interval tallies for the sample variance fell slightly below the expected 95 percent. This was most likely due to inconsistent performance of sample variance formulas (in Table 2). For example, estimator 2 showed more variance underestimates for $m=10$ than for the larger sample sizes of m (in Fig. 5). However, the confidence interval results for this estimator were best for $m=10$, indicating the few but large variance overestimates (for $m=10$ in Fig. 5) compensated for the more numerous variance underestimates. In addition to the sample variance inconsistencies, problems in the degrees of freedom formula could have caused the slight confidence interval underestimates.

Table 6.--Percentage of 95 percent confidence intervals containing the true population mean for 500 repeated samples.

Estimator	Sample sizes		Degrees of freedom ^{a/}	Confidence intervals	
	n	m		Simulation variance	Sample variance
1	75	10	75(76)82	95.2	92.4
		20	75(76)82	95.2	93.6
		40	74(76)81	94.8	93.2
2	75	10	9(18)27	94.4	93.0
		20	21(32)52	95.2	91.0
		40	42(48)71	96.4	92.4
3	75	10	9(12)18	94.2	92.2
		20	20(23)38	94.8	92.2
		40	40(44)64	95.0	91.6

^{a/}The degrees of freedom (df) are for the sample variance confidence intervals and are computed from Cochran's (1977, p. 96) formula. Given are the minimum, median (in brackets), and maximum df from each frequency distribution of 500. The simulation variance confidence intervals required no df variable because they were from the normal distribution.

Conclusions

This study concludes with two questions:

- (1) Which of the three estimators is best?
- (2) Is the error from a volume equation great enough to warrant consideration for a timber survey?

Estimator 1 exhibited the least bias. But no estimator exceeded a 1 percent bias, making bias a moot point in choosing among estimators. Comparison of the variances showed estimator 1 had the smallest variance, particularly when the phase-two sample for estimators 2 and 3 was m=10. However, when variances were used to compute confidence intervals, there was no clear distinction among estimators. All simulation variance confidence intervals performed close to the expected 95 percent probability level. The sample variance confidence intervals were all 1 to 4 percent below the 95 percent probability level.

Because neither statistical bias nor variance evaluations clearly separated the three estimators, the choice for best can be based upon other considerations. Foresters have economic and practical reasons to take the fewest number of volume measurement plots possible. Therefore, the flexibility for subsampling volume measurement clusters in phase two makes estimators 2 and 3 much more appealing than estimator 1. Furthermore, estimator 3 seems preferable over estimator

2 because it is easier to compute. Estimator 3 only requires regression parameters estimated once from all volume measurements combined, instead of parameter estimates for each cluster selected in phase two.

An underlying assumption in this study was the importance of adding volume equation error to the overall sampling error. The assumption was examined by using estimator 3 and partition 1 of its sample variance formula (in Table 2) as a way to ignore volume equation error and still obtain a sample mean and a sample variance in a timber survey. Dropping terms from a sample variance has no effect on estimator bias but does affect estimator precision. For estimator 3, the lack of partition 3 (in Table 2) in its sample variance resulted in sample variance reductions of 18, 13, and 7 percent for respective phase-two sample sizes m=10, m=20, and m=40. The reduced sample variance for estimator 3 was then used to recompute its 95 percent confidence intervals given in Table 6. The results in percentages were 90.4, 90.8, and 91.2 for respective sample sizes m=10, m=20, and m=40. Comparing these to the estimator 3 confidence intervals in Table 6 indicated a small drop in percentage of correct intervals due to ignoring partition 3.

From the reduced variance analyses for estimator 3, volume equation error is most important for the phase-two sample sizes m=10 and m=20. This implies that volume equation error need only be considered if the phase-one to phase-two sampling fraction is less than about 25 percent. Because two-phase sampling fractions less than 25 percent are probably most desirable in practice, volume equation error is likely to be of some significance in timber volume surveys. Therefore, estimator 3 or some similar estimator should be used in timber surveys to account for volume equation error.

References

- Abramowitz, M.; Stegun, I. A. Handbook of mathematical functions with formulas, graphs and mathematical tables. Applied Mathematics Series 55. Washington, DC: U.S. Department of Commerce, National Bureau of Standards; 1964. 1046 p.
- Bose, C. The variance of the forecasted mean value subjecting two-way fluctuations. Science and Culture. 7(10): 514; 1941-42.
- Chojnacky, D. C. Model-based two-phase cluster sampling of tree volume. Ph.D. dissertation. Fort Collins, CO: Colorado State University; 1985. 135 p.
- Cochran, W. G. Sampling techniques. 3d ed. New York: John Wiley and Sons; 1977. 428 p.
- Cunia, T.; Michelakackis, J. On the error of tree biomass tables constructed by two-phase sampling design. Canadian Journal of Forestry Research. 13: 303-313; 1983.

Kempthorne, O.; Folks, L. Probability, statistics and data analysis. Ames, IA: The Iowa State University Press, 1971. 555 p.

Khan, S.; Tripathi, T. P. The use of multivariate auxiliary information in double sampling. Journal of Indian Statistical Association. 5: 42-48; 1967.

Pfefferman, D.; Nathan, G. Regression analysis of data from complex samples. Bulletin of the International Statistical Institute. 47(3): 21-42; 1977.

LIST OF SYMBOLS USED IN TABLES 1 AND 2

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_i \bar{x}_i$$

$$\hat{\beta} = \frac{\sum_{i=1}^m \sum_{j=1}^{t_i} (x_{ij} - \bar{x}) y_{ij}}{\sum_{i=1}^m \sum_{j=1}^{t_i} (x_{ij} - \bar{x})^2}$$

$$\hat{\beta}_i = \frac{\sum_{j=1}^{t_i} (x_{ij} - \bar{x}_i) y_{ij}}{\sum_{j=1}^{t_i} (x_{ij} - \bar{x}_i)^2}$$

$$d_i = (\hat{\alpha}_i T_i + \hat{\beta}_i X_i - R X_i) (\bar{X}/\bar{X})$$

$$\bar{d} = \sum_{i=1}^m \frac{d_i}{m}$$

D²H = d.b.h. squared times total tree height

m = number of clusters drawn in phase two

n = number of clusters drawn in phase one

N = total number of clusters in the population

$$R = \hat{Y} / \bar{X}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^m \sum_{j=1}^{t_i} (y_{ij} - \hat{\alpha} - \hat{\beta} x_{ij})^2}{(mt - 2)}$$

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{t_i} (y_{ij} - \hat{\alpha}_i - \hat{\beta}_i x_{ij})^2}{(t_i - 2)}$$

$$s_d^2 = \frac{\sum_{i=1}^m (d_i - \bar{d})^2}{(m-1)}$$

$$s_x^2 = \frac{\sum_{i=1}^m \sum_{j=1}^{t_i} (x_{ij} - \bar{x})^2}{(mt - 1)}$$

$$s_{x_i}^2 = \sum_{j=1}^{t_i} \frac{(x_{ij} - \bar{x}_i)^2}{(t_i - 1)}$$

$$s_{Y_1}^2 = \sum_{i=1}^n \frac{(\hat{Y}_{1i} - \bar{Y}_1)^2}{(n-1)}$$

$$s_{Y_2}^2 = \sum_{i=1}^m \frac{(\hat{Y}_{2i} - \bar{Y}_2)^2}{(m-1)}$$

$$s_{Y_3}^2 = \sum_{i=1}^n \frac{(\hat{Y}_{3i} - \bar{Y}_3)^2}{(n-1)}$$

t_i = number of trees sampled in a cluster

T_i = number of trees per cluster

$$\bar{t} = \sum_{i=1}^m \frac{t_i}{m}$$

$$\bar{T} = \sum_{i=1}^m \frac{T_i}{m}$$

$$\bar{T}' = \sum_{i=1}^n \frac{T_i}{n}$$

v = sample variance

x_{ij} = D²H for jth tree of ith cluster

$$\bar{x} = \frac{\sum_{i=1}^m \sum_{j=1}^{t_i} x_{ij}}{\sum_{i=1}^m t_i}$$

$$\bar{x}_i = \sum_{j=1}^{t_i} \frac{x_{ij}}{t_i}$$

$$X_i = \sum_{j=1}^{T_i} x_{ij}$$

$$\bar{X}_i = \sum_{j=1}^{T_i} \frac{x_{ij}}{T_i}$$

$$\bar{X} = \frac{\sum_{i=1}^m X_i}{\sum_{i=1}^m T_i}$$

$$\bar{X}' = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n T_i}$$

$$\bar{X} = \sum_{i=1}^m \frac{X_i}{m}$$

$$\bar{X}' = \sum_{i=1}^n \frac{X_i}{n}$$

y_{ij} = observation for jth element of ith cluster (volume of a tree)

$$\bar{y} = \frac{\sum_{i=1}^m \sum_{j=1}^{t_i} y_{ij}}{\sum_{i=1}^m t_i}$$

$$\bar{y}_i = \sum_{j=1}^{t_i} \frac{y_{ij}}{t_i}$$

$$\hat{Y}_{1i} = \hat{\alpha}_i T_i + \hat{\beta}_i X_i$$

$$\hat{Y}_{2i} = (\hat{\alpha}_i T_i + \hat{\beta}_i X_i) (\bar{X}/\bar{X})$$

$$\hat{Y}_{3i} = \hat{\alpha} T_i + \hat{\beta} X_i$$

$$\bar{Y}_{..} = \sum_{i=1}^N \sum_{j=1}^{T_i} \frac{y_{ij}}{N}$$

$$\hat{\bar{Y}}_1 = \text{see table 1}$$

$$\hat{\bar{Y}}_2 = \text{see table 1}$$

$$\hat{\bar{Y}}_3 = \text{see table 1}$$

$$\hat{\bar{Y}} = \frac{1}{m} \sum_{i=1}^m \hat{\alpha}_i T_i + \hat{\beta}_i X_i$$

"HIGH ORDER REGRESSION MODELS FOR REGIONAL
VOLUME EQUATIONS"

Joe P. McClure and Raymond L. Czaplewski

The authors are respectively, Project Leader, Forest Inventory and Analysis, USDA Forest Service, Southeastern Forest Experiment Station, Asheville, NC 28804; and Mathematical Statistician, Multiresource Inventory Techniques, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO 80521.

Four regression models were compared in estimating total or merchantable volume using diameter, and total or merchantable height for loblolly pine and white oak. No model was consistently best for unbiased predictions. However, the higher order weighted quadratic and segmented models were more reliable than the simple weighted and unweighted models.

Introduction

Volume equations using diameter at breast height (D) and tree height (H), as the single independent variable D^2H , in the simple equation $V = a_0 + a_1 (D^2H)$, have a long history in forest mensuration. However, experience at the USDA Forest Service, Southeastern Forest Experiment Station, Forest Inventory and Analysis Project (SE FIA) has shown the need for higher-order models when a single volume equation is used for trees of all sizes.

Volume equation coefficients are usually estimated using simple linear regression. However, the variance in volume often increases with tree size (Cunia 1964). This has been demonstrated for loblolly pine and white oak by McClure et al. (1983). Homogeneous variance (homoscedasticity) is an important assumption in regression, even if the regression model is used strictly for prediction rather than for hypothesis testing. If the variance is greater near one extreme of the independent variable, then chance events in this region can have undue leverage on the regression solution. This could degrade accuracy in other regions, especially for small trees, for which there may be more observations but less variance.

In weighted regression, the model is transformed to produce a constant variance in the dependent variable for the full range of transformed tree sizes. The variance of tree volume has been reported to increase roughly proportional to the power function $(D^2H)^k$, for example: $k=0.5$ (Clutter et al. 1983); $k=1.5$ (McClure et al. 1983); $k=2$ (Cunia 1964). The volume equation is weighted by $(D^2H)^{-k/2}$ which results in a transformed model with homogeneous errors.

The objective of this study was to evaluate the effect of weighted regression and higher order prediction equations on the accuracy of volume estimation. Unbiased predictors of individual tree volume are important in extensive forest inventory because errors in applying such models approach zero as the number of tallied trees becomes large if the variance remains consistent.

Methods

Stem profiles of sample trees used were measured and tree volumes computed using methods described by Cost (1978). Cubic-foot volumes of individual standing trees (42% of available data) were estimated nondestructively using a McClure Mirror Caliper (McClure 1969) and marked section poles (McClure 1968) on a 5 to 10% subsample of regular SE FIA sample plots in Virginia, North Carolina, South Carolina, Georgia, and Florida. Also, felled trees were measured at hundreds of active logging operations distributed throughout these states. Both sources of data were pooled and treated as equivalent.

Trees were randomly assigned to either developmental or test data groups so that residuals used to compare models were independent of the errors in estimating regression coefficients (Reynolds 1984). The developmental data set contained 4134 loblolly pine and 984 white oak trees; the test data set contained 1000 loblolly pine and 500 white oak trees. Two measures of volume were considered: total (V_t) and merchantable (V_m). The former includes estimated cubic foot^m inside bark volume of the main stem (ground level to tree top), forks, major limbs, and minor limbs (0.5 to 4.0 inches at limb base, occurring on main stem, forks, or major limbs). Merchantable volume is restricted to inside bark mainstem material below a 4-inch top diameter outside bark (dob) and above a 1-foot stump height. Total tree height (H_t) and height to a 4-inch dob (H_4) were combined with squared diameters at breast height (D^2) to produce two independent variables predicting volume: D^2H_t and D^2H_4 .

Trees are described by diameter classes using standard Forest Service definitions. Trees between 1- 5-inches D are labeled saplings (trees smaller than 1-inch D, i.e., seedlings, were not measured). Loblolly pine trees greater than 9-inches and white oak trees greater than 11-inches D are called sawtimber. Intermediate-sized trees are called poletimber. Saplings are used only for estimating V_t using D^2H_t .

Four forms of prediction models were evaluated: simple weighted (1), simple unweighted (2), quadratic weighted (3), and segmented weighted (4).

$$V = b_0 + b_1X + \epsilon, \quad \epsilon \sim N(0, \sigma^2 X^k) \quad (1)$$

$$V = b_0 + b_1X + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \quad (2)$$

$$V = b_0 + b_1X + b_2X^2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2 X^k) \quad (3)$$

$$V = (b_0 + b_2 a_1) + (b_1 - b_2) X + \epsilon \text{ for } X < a_1,$$

$$= b_0 + b_1 X + \epsilon \text{ for } a_1 \leq X \leq a_2$$

$$= (b_0 + b_3 a_2) + (b_1 - b_3) X + \epsilon \text{ for } a_2 \leq X$$

$$\epsilon \sim N(0, \sigma^2 X^k) \quad (4)$$

where

$$V = V_t \text{ or } V_m$$

$$X = D^2H_t \text{ or } D^2H_4$$

a_1 = join-point constants in D^2H units

b_i = parameters estimated using linear regression.

The quadratic form in (3) includes $(D^2H)^2$, which is less common for published volume

equations than D , H , D^2 , or DH covariates. Model (3) was chosen to explore the possible curvilinear relationship between D^2H and V rather than the usual linear assumption. The segmented model (4) takes the form of a continuous line, broken into three segments at the points $X=a_1$, and $X=a_2$.

The parameters for the simple, unweighted model (2) were estimated using simple regression. Parameters for the weighted models (1, 3, 4) were estimated using multiple linear regression with a zero intercept. The unweighted versions of all four models, which are used to predict volume rather than estimate parameters, have a non-zero intercept.

For parameter estimation in the weighted regression models, it was necessary to first estimate k by studying the relationship between volume variance and D^2H . Trees from the developmental data set were divided into equal intervals based on D^2H_t or D^2H_4 . The minimum number of observed trees within each of these intervals is a subjective decision, and many interval lengths were explored (Tables 1 and 2). The variance of volume within each interval was calculated and was plotted against the average D^2H_t or D^2H_4 for that interval.

Table 1. Prediction equations for volume variance in loblolly pine for the model: $\text{Var}(V) = \sigma^2 (D^2H)^k$

Volume function	Minimum sample size	Minimum number of classes	$\ln(\sigma^2)$	k	R^2	MSE
$V_t = f(D^2H_t)$	15	90	-13.25	1.65	0.939	0.1715
	20	71	-12.85	1.60 ^a	.932	.1808
	25	56	-12.13	1.51	.916	.1791
	30	43	-11.32	1.41	.910	.1639
	40	30	-10.42	1.28	.899	.1629
	50	20	-9.60	1.11	.883	.1440
	75	9	-12.04	1.46	.849	.0467
$V_t = f(D^2H_4)$	15	83	-12.35	1.56	.939	.1223
	20	65	-12.13	1.53 ^a	.934	.1247
	25	56	-11.89	1.50	.929	.1294
	30	44	-11.39	1.43	.927	.1191
	40	27	-10.13	1.26	.915	.1011
	50	19	-9.04	1.10	.881	.0863
	75	8	-6.86	.77	.602	.0897
$V_m = f(D^2H_t)$	15	90	-9.30	1.20 ^a	.716	.5450
	20	71	-8.33	1.07	.653	.5952
	25	56	-8.44	1.06	.635	.5575
	30	43	-7.14	.90	.528	.6034
	40	30	-5.30	.64	.360	.6436
	50	20	-2.43	.24	.086	.5470
	75	9	-1.15	.14	.898	.0270
$V_m = f(D^2H_4)$	15	83	-14.10	1.73	.965	.0838
	20	65	-13.87	1.70 ^a	.961	.0890
	25	56	-13.74	1.69	.957	.0957
	30	44	-13.36	1.64	.961	.0795
	40	27	-12.57	1.53	.958	.0704
	50	19	-11.91	1.43	.951	.0558
	75	8	-10.63	1.23	.903	.0373

^aSubjectively selected as the best predictor of variance in volume given tree size.

The variance increased with tree size and could be reasonably predicted by the power model

$$\text{Var}(V) = \sigma^2(D^2H)^k \quad (5)$$

An example and illustration of this relationship is provided by McClure et al. (1983). Logarithmic transformation of (5) yields a linear model for which parameters were estimated using simple linear regression:

$$\ln[\text{Var}(V)] = \ln(\sigma^2) + k[\ln(D^2H)] \quad (6)$$

The results using (6) are given in Tables 1 and 2.

To estimate k, it is necessary to select a minimum sample size for D²H intervals to calculate variance in volume as a function of tree size (D²H). This variance is a function of interval length because the variance increases with the range of D²H values in any one interval. Also, there is only one data point per interval (mean D²H and Var(V) for that interval) that can be used in the regression model (6) for estimating k. These factors encourage use of short D²H intervals. However, as interval length decreases, so does the sample size for estimating variance within

that interval. Therefore, interval length is a compromise between sample size and homogeneity in tree size, which is necessarily a subjective decision. The selected estimates of k in Tables 1 and 2 yield rather smooth and well-defined relationships between Var(V) and tree size. We recommend sample sizes of 12-20 trees in each D²H interval for data sets that are comparable to ours in size and number of D²H intervals. Also, k is treated as a known constant in fitting models (1-4), which is not an accurate assumption. However, these problems of subjectivity and assumptions for k are greatly mitigated by the insensitivity of parameter estimates in volume models to errors in estimating k. McClure et al. (1983) suggest that k=1.5 is reasonable for most tree species.

The join-point parameters (a₁ and a₂) in the segmented model (4) are nonlinear. They cannot be simultaneously estimated using multiple linear regression. Multiple linear regressions using a wide range of join-point values were performed. The two join-points were iteratively varied in increments of 1,000 D²H units from 1,000 to 40,000, subject to a₁ ≤ (a₂ - 1,000). The join-points from the multiple regression with the smallest residual mean square error were selected as the best

Table 2. Prediction equations for volume variance in white oak for the model: $\text{Var}(V) = \sigma^2(D^2H)^k$

	Minimum sample size	Number of classes	$\ln(\sigma^2)$	k	R ²	MSE
$V_t = f(D^2H_t)$	15	26	-9.89	1.21	0.756	0.4498
	12	20	-8.57	1.02	.748	.3446
	13	19	-8.43	1.40 ^a	.745	.3473
	14	16	-14.46	1.78	.853	.1307
	16	12	-14.62	1.80	.861	.0914
	18	11	-15.64	1.94	.882	.0820
	20	6	-11.05	1.31	.475	.1051
$V_t = f(D^2H_4)$	10	29	-10.70	1.40	.809	.2853
	12	22	-10.88	1.43 ^a	.857	.2294
	14	16	-9.84	1.28	.809	.2204
	16	13	-11.48	1.49	.905	.0647
	18	9	-11.95	1.56	.905	.0910
	20	7	-11.18	1.44	.889	.1012
$V_m = f(D^2H_t)$	10	26	-4.52	5.15	.200	1.0104
	12	20	-2.17	1.82	.040	.7799
	14	16	-14.92	1.81	.871	.1150
	16	12	-15.26	1.86 ^a	.911	.0582
	18	11	-16.10	1.97	.925	.0514
	20	6	-16.06	1.97	.748	.0729
	$V_m = f(D^2H_4)$	10	29	-13.04	1.62	.874
12		22	-13.04	1.63	.916	.1650
14		16	-12.64	1.57	.900	.1583
16		13	-14.30	1.79 ^a	.948	.0491
18		9	-14.90	1.89	.960	.0532
20		7	-14.17	1.78	.960	.0509

^aSubjectively selected as the best predictor of variance in volume given tree size.

estimates of a_1 and a_2 . A nonlinear minimization routine was used to further refine estimates of the joint-points. However, the nonlinear step failed to substantially improve model predictions.

Independent test data were partitioned into 20 D^2H classes for loblolly pine and 10 classes for white oak. Class intervals were chosen so that each contained approximately 50 trees from the test data set, resulting in D^2H intervals of varying width. Residual error is defined as the observed volume (V) minus the predicted volume (V) using models^o(1-4). The variance of the weighted least square residual (Draper and Smith 1981) was made homogeneous for the entire D^2H range by applying the same weighting transformation used to estimate model parameters in weighted regression:

$$R_i = (V_{oi} - V_i)(D_i^2 H_i)^{-k/2}$$

The specific values of k for each of the eight combinations of species, volume type, and height type are footnoted in Tables 1 and 2.

The null hypothesis that the weighted least square residual is normally distributed given an estimated mean and variance was tested at the 0.05 significance level using both the Kolmogorov-Smirnov and Crámer-von Mises statistics (Reynolds 1984). This assumption is needed later to test the hypothesis that the mean of the residuals equals zero. If this Gaussian hypothesis of normally distributed residuals was rejected using one statistic but not rejected using the other, then the null hypothesis was rejected for that case. This Gaussian hypothesis was tested for each of the 32 combinations of species, independent and dependent variables, and model form. There were 973 to 1,000 loblolly pine and 481 to 500 white oak trees used in each such test (sapling size trees were not used if the model included V_m or H_4). For each rejected Gaussian hypothesis, similar tests were performed for each of the 10 to 20 D^2H classes so that Gaussian methods could be validly applied to portions of the D^2H range. There were 359 such tests, most of which used 45-55 trees.

For each D^2H class for which the Gaussian hypothesis was not rejected, a test for bias was performed using the weighted least square residuals. The null hypothesis is that the mean of the residual error for all trees in the D^2H class interval is zero (i.e., unbiased). This hypotheses was tested by constructing a 95% confidence interval for the mean weighted least square residual (\bar{R}):

$$\bar{R} \pm t5(n^{-1/2})$$

where

$$S^2 = \sum_{i=1}^n (R_i - \bar{R})^2 / n - 1$$

and t is the 97.5 percentile of the t distribution with $n-1$ degrees of freedom. The null hypothesis was accepted if this confidence interval contained zero. There were 464 such hypotheses tested (as many as 20 D^2H classes for each of the 16 loblolly pine models and 10 for each of the 16 white oak models). These tests used independent test data, and most tests used 45-55 trees.

Results and Discussion

Parameter estimates for all models are given in Table 3. The Gaussian hypothesis for the weighted least square residuals across the entire D^2H range was not rejected in only 8 cases of the 32 tests: all three weighted models for loblolly pine V_m using H_4 ; the four weighted models for white oak V_m using the simple (1) or quadratic (3) model^m forms with both H_2 and H_4 ; and the simple, unweighted white oak model (2) for V_m . The null hypothesis that the mean weighted squared error across the entire D^2H range is zero (unbiased) is not rejected in 7 of the 8 cases; the exception is the unweighted simple linear regression model (2) for oak. The large number of trees (973, 481) in the tests for goodness-of-fit and unbiasedness makes the possibility of Type II error small. For the 24 remaining combinations of species, model form, volume type, and height type, the Gaussian hypothesis was tested within each D^2H class (Table 4). When the Gaussian hypothesis was rejected, which occurred in 30 of the 359 individual D^2H classes, no tests for significant bias were performed.

The final set of hypotheses tested were those associated with unbiased predictions of volume using D^2H . These were performed for each D^2H class for which the hypothesis of normally distributed, weighted least square residuals was not rejected. The results of these tests, the magnitude and direction of any bias, and the standard deviation for each D^2H interval (S_i) of the weighted least square residuals are also given in Table 4. (S_i was essentially the same for all four model forms.) These statistics can be used for computing prediction, tolerance, and confidence intervals (Reynolds 1984). The confidence interval was used to test the hypothesis that the mean of the weighted least square residual is zero (i.e., unbiased estimate of volume). It is these tests that directly address the objective of this study.

The unweighted model form (2) has a strong bias (overestimated volume) for saplings and small poletimber of both species using the D^2H_4 covariate (Table 4). Although these smaller trees are not as commercially valuable as larger trees, they often represent a major portion of the standing volume of an inventory unit, especially in commercially active geographic areas. However, the unweighted model (2) was significantly less biased than the weighted models (1, 3, 4) for these small

trees when D^2H_t is used to estimate V_m rather than D^2H_4 . Conversely, V_t estimates for white oak saplings were less biased using the weighted models (1, 3, 4).

All models usually produced unbiased estimates for large poletimber and all but the largest sawtimber classes (Table 4). All models yielded biased estimates for saplings--overestimated for loblolly pine, and underestimated for oak. Most of the differences in bias among models were for small- and mid-sized poletimber and very large sawtimber. The magnitude of bias for smaller trees was much less for weighted regression models (1, 3, 4) than for unweighted model (2) using D^2H_4 . To a lesser extent, the opposite trend was observed for the D^2H_t is covariate.

There was no consistently best model formulation for unbiased predictions of volume for small- and mid-sized poletimber. The weighted simple model (1) was less biased in estimating V_t for oak using D^2H_4 , while there was less bias using the unweighted simple model (2) in predicting V_m for oak or V_t for loblolly pine using D^2H_4 (Table 4). There was little difference between (1) and (2) in all other cases. The weighted quadratic model (3) was somewhat better than the weighted simple model

(2) for estimating V_t using D^2H_4 , but there was little difference otherwise. The weighted segmented model (4) was the most consistent formulation for unbiased estimates of loblolly pine volume. However, the quadratic model (3) also did well for loblolly pine, and it slightly out-performed the segmented formulation for white oak trees in this size range.

For very large sawtimber, the segmented model (4) was usually best for loblolly pine (Table 4). The weighted simple model (1) tended to produce biased estimates of white oak volume for such trees, while the other models were unbiased for oak. There was little difference among models for smaller sawtimber. The least number of biased D^2H classes in loblolly pine was obtained using the segmented model (4), while the quadratic model (3) was the next best. For white oak, the quadratic model (3) was usually best for sawtimber, with the segmented (4) and unweighted simple (2) models being the next best.

The variance of weighted residuals should be relatively constant for all tree sizes. This is generally true in Table 4, with three exceptions: loblolly pine $V_m = f(D^2H_t)$, and white oak $V_t = f(D^2H_4)$ and $f(D^2H_t)$. There is a trend for the variance to increase as tree size

Table 3. Estimated Parameters for Volume Equations^a

	b_0	b_1	b_2	b_3	a_1	a_2
<u>LOBLOLLY PINE</u>						
$V_m = f(D^2H_4), k=1.70, n=4002$						
Simple Weighted (1)	0.4257	0.002528				
Simple Unweighted (2)	1.3419	0.002367				
Quadratic Weighted (3)	0.3690	0.002612	-6.397×10^{-9}			
Segmented Weighted (4)	0.6213	0.002524	-4.486×10^{-4}	-1.301×10^{-4}	1,000	7,000
$V_t = f(D^2H_4), k = 1.53, n=4002$						
Simple Weighted (1)	1.1545	0.002794				
Simple Unweighted (2)	1.8217	0.002683				
Quadratic Weighted (3)	1.1070	0.002856	-4.234×10^{-9}			
Segmented Weighted (4)	1.4030	0.002157	-5.483×10^{-4}	-2.025×10^{-4}	1,000	2,000
$V_m = f(D^2H_t), k=1.20, n=4002$						
Simple Weighted (1)	-0.1886	0.002028				
Simple Unweighted (2)	-0.4329	0.002066				
Quadratic Weighted (3)	-0.1742	0.001994	1.826×10^{-9}			
Segmented Weighted (4)	-0.8456	0.002117	7.946×10^{-4}	-1.420×10^{-4}	1,000	2,000
$V_t = f(D^2H_t), k=1.60, n=4132$						
Simple Weighted (1)	0.0187	0.002298				
Simple Unweighted (2)	-0.01871	0.002343				
Quadratic Weighted (3)	0.0217	0.002266	2.454×10^{-9}			
Segmented Weighted (4)	-0.1413	0.002304	1.712×10^{-4}	0.641×10^{-4}	1,000	6,000

Table 3. Continued

	b_0	b_1	b_2	b_3	a_1	a_2
<u>WHITE OAK</u>						
$V_m = f(D^2H_d), k=1.79, n=959$						
Simple Weighted (1)	0.6054	0.002695				
Simple Unweighted (2)	1.6202	0.002496				
Quadratic Weighted (3)	0.5363	0.002799	-8.085×10^{-9}			
Segmented Weighted (4)	0.8141	0.002666	-3.989×10^{-4}	-2.158×10^{-4}	1,000	12,000
$V_t = f(D^2H_t), k=1.43, n=959$						
Simple Weighted (1)	1.4280	0.003243				
Simple Unweighted (2)	12.1424	0.003127				
Quadratic Weighted (3)	1.3550	0.003327	-4.842×10^{-9}			
Segmented Weighted (4)	1.9400	0.003190	-2.174×10^{-4}	-2.326×10^{-4}	3,000	27,000
$V_m = f(D^2H_t), k=1.86, n=959$						
Simple Weighted (1)	-0.0491	0.001886				
Simple Unweighted (2)	-0.2660	0.001970				
Quadratic Weighted (3)	-0.0467	0.001829	4.693×10^{-9}			
Segmented Weighted (4)	1.2613	0.002445	12.330×10^{-4}	-4.657×10^{-4}	1,000	2,000
$V_t = f(D^2H_t), k=1.40, n=984$						
Simple Weighted (1)	0.0186	0.002439				
Simple Weighted (2)	-0.2596	0.002473				
Quadratic Weighted (3)	0.0223	0.002420	1.061×10^{-9}			
Segmented Weighted (4)	-0.3225	0.002458	0.345×10^{-4}	-0.653×10^{-4}	1,000	25,000

^aUnits are: v--cubic feet; d--inches; H_t, H_d --feet

increases in these three cases, suggesting that the corresponding estimate of k is too small.

In this study, a large number of trees were randomly divided into developmental and test groups. It is expected that models fit to one such group can be applied to the second with little bias. Our results merely compared several model forms as an approximation of the true relationship between D^2H and volume. If simple random sampling of trees from a large geographic region is conducted, then our results would apply directly to such samples.

Conclusions

There was little difference among the four models studied in estimating volume of commonly available, commercial sawtimber. However, there were differences among models in predicting pole timber volume when a single equation was used for trees in all size classes. Unfortunately, there was no obviously best model for producing unbiased predictions. Even simple, unweighted regression performed very well in some cases. Higher order weighted models, such as the quadratic and segmented, tended to be less biased and more dependable than simpler models.

Literature Cited

- Clutter, J. L., J. C. Fortson, L. V. Pienar, G. H. Brister, and R. L. Bailey. 1983. Timber management: A quantitative approach. John Wiley and Sons. N.Y.
- Cost, Noel D. 1978. Multiresource inventories: A technique for measuring volume in standing trees. USDA For. Serv. Res. Paper SE-196. 18 p.
- Cunia, T. 1964. Weighted least squares method and construction of volume tables. For. Sci. 10:180-191.
- Draper, N. R. and H. Smith. 1981. Applied Regression Analysis. John Wiley and Sons N.Y. 709 pp.
- McClure J. P. 1968. Sectional aluminum poles improve length measurements in standing trees. USDA For. Serv. Res. Note SE-98. 4 p.
- McClure, J. P. 1969. The minor caliper, or new optical dendrometer. USDA For. Serv. Res. Note SE-112. 5 p.

McClure, Joe P., Hans T. Schreuder, and Rodney L. Wilson. 1983. A comparison of several volume tables equations for loblolly pine and white oak. USDA For. Serv. Res. Paper SE-240. 8 p.

Reynolds, M. R. 1984. Estimating the error in model predictions. For. Sci. 30(2):454-459.

Table 4. Means and Standard Deviation of Weighted Least Square Residuals by D²H Class: $R = (V_o - V) / (D^2H)^{-k/2}$

Midpoint D ² H class	Mean for each model form ^a				Var(R) ^b	Test sample size (n)	
	1	2	3	4			
<u>LOBLOLLY PINE</u>							
<u>$V_m = f(D^2H_4)$ k=1.70</u>							
Poletimber	350	-8.39 ^d	-69.33 ^c	-6.35 ^d	-11.32 ^d	10.83	45
	530	2.78 ^d	-32.50 ^c	3.13 ^d	6.19 ^d	7.57	54
	880	2.77 ^d	-21.69 ^d	2.39 ^d	9.10 ^d	8.06	49
	1,130	2.43 ^d	-16.45 ^d	1.68	-2.46	10.04	40
	1,420	3.52 ^d	-7.22 ^d	2.26 ^d	0.29	9.83	59
	2,350	1.51	-5.95 ^d	0.08	-1.05	8.24	45
Sawtimber	2,900	2.07	-3.04 ^c	0.55	-0.02	9.27	57
	3,500	2.57 ^d	-0.73	1.03	0.83	9.56	71
	4,400	2.36	0.69	0.85	0.93	9.05	45
	5,300	3.40 ^d	2.97 ^d	1.98	2.21	8.59	55
	6,300	-0.49	0.02	-1.79	-1.51	9.18	47
	7,350	-1.29	0.08 ^c	-2.40	1.77	10.47	48
	8,550	4.43 ^d	6.53 ^d	3.55 ^d	8.75 ^d	9.76	44
	9,950	-0.33	2.44	-0.89	4.23 ^d	9.61	43
	11,500	-1.62	1.75	-1.78	3.16	11.83	45
	13,550	-2.53 ^d	1.39	-2.19	2.46	10.80	42
	15,400	-3.82 ^d	0.59	-2.89	1.37	10.82	36
	21,500	-2.81	2.28	-0.68	2.63	13.29	42
	50,000	-8.55 ^d	-2.53	-3.54 ^d	-2.61	10.59	55
<u>$V_t = f(D^2H_4)$, k=1.53</u>							
Poletimber	350	-18.11 ^d	-90.96 ^d	-15.01 ^d	-23.65 ^d	40.31	45
	630	6.38 ^d	-38.19 ^d	7.22 ^d	14.33 ^d	24.45	54
	880	3.80 ^d	-28.19 ^d	3.60	18.63 ^d	21.69	49
	1,130	2.92	-22.37 ^d	2.15	-8.72	20.84	40
	1,420	10.35	-9.27 ^d	9.10 ^d	2.80	19.06	51
	2,350	-0.08	-10.97 ^d	-2.04	-4.41	18.67	45
Sawtimber	2,800	3.73	-4.01	1.56	0.56	18.25	57
	3,500	5.79 ^d	0.52	3.50	3.63	23.23	71
	4,400	6.26	3.33	3.93	4.86	22.42	45
	5,300	6.18	5.06	3.89	5.44	20.13	55
	6,300	-1.98 ^c	-1.68 ^c	-4.16 ^c	-2.21 ^c	20.76	47
	7,450	-2.13 ^c	-0.50 ^c	-4.11 ^c	-1.88 ^c	23.95	48
	8,550	10.95 ^d	13.72 ^d	9.26 ^d	11.62 ^d	23.00	44
	9,950	-0.45	3.42	-1.73	0.50	23.53	43
	11,500	-2.51	2.38	-3.25	-1.66	28.53	45
	13,650	-3.44	2.40	-3.47	-1.67	27.59	42
	16,400	-6.43	0.28	-5.59	-4.35	25.86	36
	20,500	-3.33	4.66	-0.65	-0.79	32.04	42
	50,000	-17.87 ^d	-8.07 ^d	-10.66 ^d	7.93	28.32	55

Table 4. Continued

Midpoint D ² H class	Mean for each model form ^a				Var(R) ^b	Test sample size (n)	
	1	2	3	4			
<u>V_m = f(D²H_t), k=1.20</u>							
Poletimber	630	-35.84 ^d	8.91 ^d	-34.38 ^d	-30.22 ^d	23.77	5
	880	-49.34 ^d	-13.24 ^d	-46.92 ^d	-70.04 ^d	65.00	17
	1,130	-62.30 ^d	-32.06 ^d	-59.18 ^d	21.18 ^d	61.62	36
	1,420	-56.35 ^d	-31.81 ^d	-52.49 ^d	12.06	79.51	53
	1,850	-57.62 ^d	-38.33 ^d	-53.02 ^d	-3.07	63.50	66
	2,350	-38.46 ^d	-23.78 ^d	-33.20 ^d	3.95	65.61	62
	2,900	-38.88 ^d	-27.45 ^d	-33.15 ^d	-4.96	67.35	61
	3,600	-18.93	-10.65	-12.78	6.78	79.87	55
Sawtimber	4,400	-16.62	-11.69	-10.11	0.42	78.98	69
	5,300	-19.12	-16.59	-12.45	-8.24	78.59	63
	7,350	22.71 ^d	21.05 ^d	29.26 ^d	22.85 ^c	75.20	55
	8,550	-14.44	-18.09	-8.23	-19.32	77.16	46
	9,950	16.83	11.53 ^d	22.55 ^d	7.82	105.11	45
	11,600	27.68 ^d	20.70 ^d	32.65 ^d	14.46	109.50	54
	13,650	39.34 ^d	30.35	42.99	21.10 ^d	111.59	49
	16,400	24.89	13.73	26.45	1.28	116.29	57
	20,500	18.99	5.41	17.15	-10.57	144.36	47
	50,000	-9.00	-27.73	-24.50	8.81	155.43	85
<u>V_t = f(D²H_t), k=1.60</u>							
Saplings	350	2.25 ^c	112.81 ^d	1.32 ^c	84.87 ^d	14.43	11
	630	-4.36	5.56	-3.42	-1.87	10.10	12
Poletimber	880	-4.17	3.12	-3.15	-4.05	16.56	24
	1,130	-4.24	1.53	-3.16	1.39	14.69	38
	1,420	-3.91	0.42	-2.78	0.66	17.59	53
	1,850	-5.93 ^d	-2.85	-4.77 ^d	-2.26	13.93	66
	2,350	-2.72	-0.67	-1.55	0.19	13.83	62
	2,900	-4.28 ^c	-2.91 ^c	-3.11 ^c	-1.83 ^c	16.12	61
	3,600	-2.65	-1.90	-1.50	-0.63	16.19	55
Sawtimber	4,400	-1.80	-1.67	-0.71	-0.19	15.32	69
	5,300	-1.39	-1.59	-0.28	0.03	15.00	63
	6,300	2.28	1.63	3.20 ^d	0.47	17.50	48
	7,350	3.37	2.42	4.17	0.51	13.47	55
	8,550	-2.87	-4.11 ^d	-2.23	-6.03 ^d	13.43	46
	9,900	2.69	1.23	3.15	-0.69	17.53	45
	11,500	4.07	2.38	4.31	0.44	17.58	54
	13,650	5.57 ^d	3.73	5.56 ^d	1.76	18.02	49
	15,400	3.22	1.04	2.56	-0.96	18.27	57
	20,500	2.94	0.49	1.67	-1.55 ^d	20.48	47
	50,000	-1.72	-4.66 ^d	-5.33 ^d	-6.84 ^d	21.16	85
<u>WHITE OAK</u>							
<u>V_m = f(D²H₄), k=1.79</u>							
Poletimber	430	-1.32 ^d	-39.32 ^c	-0.38	-0.49	9.45	56
	1,080	3.04 ^d	-14.24 ^d	2.52 ^d	3.76 ^d	7.67	46
	1,650	2.64 ^d	-6.66 ^d	1.60	0.10	7.46	40
	2,550	4.54 ^d	-0.32	3.28 ^d	3.02 ^d	8.08	52
	3,750	1.24	-0.71	-0.06	0.38 ^c	9.71	47
Sawtimber	5,350	0.59	0.66	-0.59	0.18	7.28	49
	7,500	-1.30	0.32	-2.16	-1.36	8.80	54
	11,950	-1.07	1.66	-1.34	0.14 ^d	8.59	56
	17,350	-2.40 ^d	1.34	-1.48	3.96 ^d	8.39	46
	40,000	-4.07 ^d	0.90	1.18	3.00	8.79	35

Table 4. Continued

	Midpoint D^2H class	Mean for each model form ^a				Var(R) ^b	Test sample size (n)
		1	2	3	4		
$V_t = f(D^2H_t), k=1.43$							
Poletimber	430	-0.93	-82.16 ^d	3.27 ^d	-48.15 ^d	42.64	56
	1,080	9.34	-34.28 ^d	8.99 ^d	-8.54	43.30	46
	1,650	10.02	-16.68 ^d	7.46 ^d	6.29 ^d	36.89	40
	2,550	20.45 ^d	4.29	16.54 ^d	25.82 ^d	44.25	52
	3,750	5.43	-3.04	0.66	-3.75 ^c	56.30	47
Sawtimber	5,350	3.75	1.24	-1.34	-1.49	42.90	49
	7,500	-4.08	-1.40	-8.86	-6.00	55.49	54
	11,960	-6.38	0.51	-9.88	-6.58	57.21	56
	17,350	-12.67	-1.27	-12.60	-9.28	62.27	46
	40,000	-11.89	6.31	5.11	26.30 ^c	88.58	35
$V_m = f(D^2H_t), k=1.86$							
Poletimber	1,080	-4.32 ^d	-2.43	-3.51 ^d	0.85	6.53	25
	1,650	-4.16 ^d	-3.29 ^d	-3.35 ^d	-0.82	5.17	55
	2,550	-1.06	-1.02	-0.30	5.66 ^d	4.81	48
	3,750	-0.35	-0.09	1.04	4.50 ^d	4.19	53
Sawtimber	5,350	0.09	-0.67	0.67	2.65 ^c	5.13	50
	7,500	0.00	-1.01	0.40	1.31	4.82	52
	10,950	0.44	-0.78	0.54	0.79	4.88	73
	17,350	1.78 ^d	0.38	1.34 ^d	1.38 ^d	4.50	66
	40,000	2.36 ^d	0.75	0.01	1.19	4.80	58
$V_t = f(D^2H_t), k=1.40$							
Saplings	430	13.82 ^c	191.55 ^c	12.20 ^c	231.25 ^c	22.65	19
Poletimber	1,080	4.47	22.76 ^d	5.64	25.94 ^d	34.54	26
	1,650	-7.32	5.45 ^c	-5.95	7.26 ^c	30.46	55
	2,550	-1.71	6.40	-0.15	6.87	32.42	48
	3,750	1.03	5.85	2.69	5.55	39.91	53
	5,350	-6.61 ^d	-4.06	-4.95	-5.02	43.55	50
Sawtimber	7,500	-12.74 ^d	-12.23	-11.21	-13.82 ^d	44.09	52
	10,950	-10.99	-12.38 ^d	-9.86	-10.17	53.05	73
	17,350	-3.21	-6.39	-3.05	-2.92	55.15	66
	40,000	10.71	4.81	5.97	20.73 ^c	67.31	58

^aRegression model forms are: (1) weighted simple; (2) unweighted simple; (3) weighted quadratic; and (3) weighted segmented.

^bVariance for all four models was identical, only the mean error varied between models. ^cHypotheses of normally distributed least square residual rejected.

^dDo not reject hypothesis of normally distributed least square residual, but reject hypothesis mean weighted residual equals zero. For those means not footnoted, both the Gaussian and unbiased hypotheses were not rejected.

Wharton, Eric H.; Cunia, Tiberius. 1987. Estimating tree biomass regressions and their error, proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates; 1986 May 26-30; Syracuse, New York. NE-GTR-117. Broomall, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 303 p.

Proceedings of a workshop co-sponsored by the USDA Forest Service, the State University of New York, and the Society of American Foresters. Presented were papers on the methodology of sample tree selection, tree biomass measurement, construction of biomass tables and estimation of their error, and combining the error of biomass tables with that of the sample plots or points. Also presented were papers on various aspects of biomass research currently being conducted in the United States, Canada, and abroad.

Keywords: Biomass, regression functions, regression error, sampling error, measurement error.

Headquarters of the Northeastern Forest Experiment Station are in Broomall, Pa.
Field laboratories are maintained at:

- Amherst, Massachusetts, in cooperation with the University of Massachusetts.
- Berea, Kentucky, in cooperation with Berea College.
- Burlington, Vermont, in cooperation with the University of Vermont.
- Delaware, Ohio.
- Durham, New Hampshire, in cooperation with the University of New Hampshire.
- Hamden, Connecticut, in cooperation with Yale University.
- Morgantown, West Virginia, in cooperation with West Virginia University, Morgantown.
- Orono, Maine, in cooperation with the University of Maine, Orono.
- Parsons, West Virginia.
- Princeton, West Virginia.
- Syracuse, New York, in cooperation with the State University of New York College of Environmental Sciences and Forestry at Syracuse University, Syracuse.
- University Park, Pennsylvania, in cooperation with the Pennsylvania State University.
- Warren, Pennsylvania.

Persons of any race, color, national origin, sex, age, religion, or with any handicapping condition are welcome to use and enjoy all facilities, programs, and services of the USDA. Discrimination in any form is strictly against agency policy, and should be reported to the Secretary of Agriculture, Washington, DC 20250.