



United States
Department of
Agriculture

Forest Service

Northeastern Forest
Experiment Station

NE-GTR-117



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

PROGRAM PLANNING COMMITTEE

Alexander Clark III
USDA Forest Service
Southeastern Forest
Experiment Station
Athens, Georgia

Vernon J. LaBau
USDA Forest Service
Pacific Northwest Forest and
Range Experiment Station
Anchorage, Alaska

Tiberius Cunia
State University of New York
College of Environmental
Science and Forestry
Syracuse, New York

Eric H. Wharton
USDA Forest Service
Northeastern Forest
Experiment Station
Broomall, Pennsylvania

This material was prepared with the support of the U.S. Department of Energy Grant No. 23-524, which was administered by the U.S. Department of Agriculture, Forest Service. Each author is responsible for the accuracy and style of his own paper. Any opinions, findings, conclusions, or recommendations of those authors outside the USDA Forest Service do not necessarily reflect the policy of the U.S. Department of Energy or the USDA Forest Service.

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**

**May 26-30, 1986
Syracuse, New York**

Sponsored by:

State University of New York
College of Environmental Science and Forestry

USDA Forest Service
Northeastern Forest Experiment Station

Society of American Foresters
Forest Inventory Working Group

Compiled by:

Eric H. Wharton
Research Forester
USDA Forest Service
Northeastern Forest Experiment Station
Broomall, Pennsylvania

Tiberius Cunia
Professor of Statistics and Operations Research
State University of New York
College of Environmental Science and Forestry
Syracuse, New York

PREFACE

It is common to calculate forest inventory estimates and their error by procedures which ignore the error of tree volume or biomass tables. For this reason, the error of forest inventory estimates as calculated do not usually include errors due to table development. This technique has become acceptable for volume estimates because the age of volume tables prevent calculating errors associated with table development, but this argument does not apply to biomass tables. The sample data from which biomass tables were constructed are still available and can be used to estimate the error of biomass regression functions from which the tables were constructed.

Procedures exist to include the error of biomass tables when the error of the biomass inventory estimates are calculated, but they are not readily available. This initiated the present workshop jointly sponsored by the State University of New York, the USDA Forest Service, and the Society of American Foresters: held in Syracuse, New York, May 26-30, 1986.

The error associated with table development, be it volume, biomass, or some other measure, may have a significant influence on the total sampling error of an estimate. But how much does the error of table development contribute to the total error of the estimates? In some cases, it may have more of an influence than the error associated with field plots or points that are used to collect resource information. The degree of influence is not always known, and not including the error associated with table development along with errors that result from sample plots or points may be costly.

It is the hope of the planning committee to present techniques whereby the errors associated with table development, specifically biomass tables, may be combined with the error associated with sample plots used in resource inventories. With this knowledge, reasonably sound judgements can be made about the validity of the inferences made concerning the error of the forest biomass inventory.

To this end, the primary topics discussed at the workshop included methods for sample tree selection, tree biomass measurement, the construction of biomass tables, the estimation of the error associated with biomass tables, and combining the error of biomass tables with those of sample plots. These topics were covered during lecture and discussion sessions conducted by Tiberius Cunia, Faculty of Forestry, State University of New York, College of Environmental Science and Forestry.

Because most of the material contained in the lecture sessions is either not published, or if published is in the form of research papers, we have included ten papers covering the lecture notes in Part I of the proceedings. The first section (two papers) introduces a procedure for combining the error of sample plots with that of biomass regressions. The second section (three papers) shows how to apply the weighted least squares method and calculate the required statistics of the biomass regressions when (1) the sample trees are assumed to be selected by random sampling, cluster sampling, and double sampling; and (2) some of the independent variables are allowed to be qualitative, such as when sample trees are classified into several groups. The third section (five papers) show how to calculate the required statistics of the sample plot data so that the error of the sample plots would be expressed in a form suitable for combination with the error of biomass regression.

We have considered the case of sample plots selected by simple random sampling, stratified sampling, double sampling for stratification, two-stage sampling, and double sampling with regression estimators. We have considered estimates of average biomass and biomass growth per acre calculated from Continuous Forest Inventory Systems with or without Sampling with Partial Replacement (SPR).

Also discussed were various aspects of biomass research that is being conducted in the United States, Canada, and abroad. The workshop provided a forum for researchers to present many different aspects of biomass research that is currently being conducted. Papers that were submitted for publication within these proceedings, but not presented during the workshop, are footnoted as contributed papers.

The research papers are presented in Part II of the proceedings. These papers included discussions of (1) measurement error in biomass table construction, (2) sampling the biomass of understory vegetation, (3) biomass regression functions and their application to the forest inventories of the eastern United States, (4) biomass studies outside the United States, and (5) the use of simulation techniques to evaluate the validity of inferences made concerning the error of the biomass regressions.

The planning committee would like to express its sincere thanks to the workshop participants for their time, knowledge, and experience that they provided, all of which greatly contributed to the success of the workshop.

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	
<u>Biomass Functions in the Eastern United States: Regression Models and Application to Timber Inventories</u>	
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	
<u>Biomass Studies Outside the United States</u>	
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	
<u>Use of Simulation Techniques to Evaluate the Validity of Biomass Regression Functions</u>	
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.

TUTORIAL PAPERS

**Combining the Error
of Sample Plots
and Biomass Regressions**

ERROR OF FOREST INVENTORY ESTIMATES: ITS MAIN

COMPONENTS^{1/}

Tiberius Cunia

Professor of Statistics and Operations Research
SUNY College of Environmental Science and Forestry,
Syracuse, NY, 13210

Most sampling designs of forest inventory consist of two major phases; a first phase where the trees are measured for diameter alone and a second phase where the trees are measured for biomass in addition to diameter. A biomass regression function estimated from the second phase sample is applied to the tree data of the first phase to estimate the average biomass per unit area. A technique is shown whereby the errors of the first and second phase samples are combined when the error of the average biomass per unit area is calculated. This technique is applied to the case where (i) the sampling technique of the first and second phase is simple random sampling, (ii) the samples of the two phases are statistically independent and (iii) the biomass regression function of the second phase is estimated by the weighted least squares linear regression techniques. Numerical examples are also given.

Introduction

Most sampling designs of forest inventory consist of two major phases. In the first phase, a relatively large sample of trees is selected (usually in clusters defined in terms of sample plots of fixed area or sample points) and the trees are measured, among other things, for their diameter at breast height d . These trees are not measured for their biomass y . In the second phase, a relatively small sample of trees is selected and the trees are measured for biomass in addition to diameter. These trees are used to estimate a relationship between tree diameter and biomass, usually but not necessarily expressed as the regression function of tree biomass on diameter. This relationship is then applied to the trees of the first phase sample to calculate forest inventory estimates of average biomass per unit area.

When previously constructed biomass regression functions are available, the second phase sample is no longer necessary. Implicitly then, a critical assumption is being made that the tree population for which the regression function was

calculated and the tree population being currently inventoried are very similar, if not identical. This is a big assumption to make, since it is generally true that the regression functions may vary considerably from one to the next forest area, even though the two areas may be expected to be similar.

Because of the basic structure of the sampling design the error of the forest inventory estimates has two main components. There is first the component due to the random selection of the sample units of the first phase. Successive applications of the same selection procedure to the same forest area result in different sets of sample trees and, thus, different sets of estimates. The size of this component is greatly affected by (i) the sampling design of the first phase, (ii) the sample size, (iii) the type of estimator used (for given sample data and required parameter to estimate, there are generally several estimators, each estimator having its own precision) and (iv) the inherent variation between the sample units (as determined by the geographical distribution of trees, variation between tree biomass as well as the population frame and sampling design used). The second component is associated with the sample of the second phase, more specifically with the error of the biomass regression. The size of this component is also affected by (i) the sampling design used to select the trees of the second phase, (ii) the sample size, (iii) the estimation procedure and (iv) the inherent variation of the tree biomass values about the regression function.

These two components constitute what is generally known as the sampling error. They are different from other error components, one of which, the measurement error, may become critically important with large samples. Defined as the difference between the true, conceptual value, and the recorded value of a sample unit, the measurement error has a random and a systematic part. The random part is expected to average to zero, in the long run, and can ordinarily be included in the first two error components above, even though conceptually different. The systematic part, the measurement bias, is seldom affected by the sample size, and for this reason its effect may become critical with large samples.

Additional error components can also be identified. One such component is that due to the statistical model used in defining the estimator. Changing the model will generally change the estimates. Different statisticians working with the same sample data will not necessarily arrive at the same estimates. This is particularly true in biomass tables construction where several regression functions may fit equally well the same sample tree data. However, when the statistical model used fits reasonably well the sample data, the error is generally small and can be ignored.

The objective of the present paper is to have a closer look at the two main error components, those associated with the first and second

^{1/} Paper based on a set of lecture notes "On the error of biomass estimates in forest inventories; Part 1: Its major components". Faculty of Forestry Miscellaneous Publication Number 8 (ESF 85-004). SUNY College of Environmental Science and Forestry, Syracuse, NY.

phase samples. We shall analyze their structure and we shall present a procedure to combine them into one, overall error value, provided that each component is expressed in an appropriate form. The effect of all other sources of error, as for example that due to measurement and statistical model will be ignored.

A Specific Approach to Combine the Two Error Components

The error of the estimators will be expressed here as the variance. This is more convenient to work with than standard error, mean squares or confidence limits. Furthermore, the variance can be split into additive components with each component associated exclusively with one source of error, provided certain conditions are satisfied.

To calculate forest biomass estimates, one would combine statistics of the second phase (biomass regression coefficients) with statistics of the first phase (average number of trees by size and quality classes). To calculate the error of these estimates, we shall now propose the following approach where, for convenience, we shall refer to a specific biomass estimator, w .

We shall make the following assumptions:

(1) The statistics of the first phase that enter into the calculation of the estimator w are denoted as $z_1, z_2, \dots, z_p$ and estimates $S_{z_i z_j}$, $i, j = 1, 2, \dots, p$, of their variances and covariances can be calculated from the sample data.

(2) The statistics of the second phase that enter into the calculation of the estimator w are denoted by $b_1, b_2, \dots, b_m$ and estimates $S_{b_i b_j}$, $i, j = 1, 2, \dots, m$, of their variances and covariances can be calculated from the sample data.

(3) The statistics of the first phase are statistically independent of the statistics of the second phase.

(4) The estimator w can be expressed as an explicit function of statistics z and b , say $w = f(z, b)$.

Then the variance of w can be approximately estimated by the expression, given among others by Davies (1961),

$$S_{ww} = \sum_{i=1}^p \sum_{j=1}^p \left(\frac{\partial f}{\partial z_i} \right) \left(\frac{\partial f}{\partial z_j} \right) S_{z_i z_j} + \sum_{i=1}^m \sum_{j=1}^m \left(\frac{\partial f}{\partial b_i} \right) \left(\frac{\partial f}{\partial b_j} \right) S_{b_i b_j} + \text{terms involving higher differentials}$$

Because the random variables z are statistically independent of the random variables b , the terms involving covariances $S_{z_i b_j}$, $i = 1, 2, \dots, p$ and $j = 1, 2, \dots, m$ have been left out from the formula above. Furthermore, the terms involving higher differentials are of a lower order of magnitude whenever the coefficients of variation

of z and b are relatively small, say less than 20 percent, the usual case in forest inventory; and, thus, all these terms can ordinarily be ignored.

The critical points of this approach are those of (i) expressing w as a simple function of variables z and b , and (ii) finding valid estimates of their variances and covariances. When the regression functions of the second phase are non-linear, the expression of w and that of its variance may be so complex that it may become extremely cumbersome to apply the formula above. On the other hand, when the regression functions are linear, the derivation of the formula of w , and that of its variance becomes, most of the time, relatively simple.

Let us, therefore, make the following more specific assumptions:

(1) The regression function of tree biomass y on tree characteristics $x_1, x_2, \dots, x_m$, where x_1 is usually, but not necessarily defined as identically equal to 1, is of the linear form

$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]' [x]$
where notation $[]$ was used to denote matrices and vectors, and $[]'$ to denote their transposes. Denote the estimate of $[\beta]'$ as the vector

$$[b]' = [b_1 \ b_2 \ \dots \ b_m]$$

and the estimate of the covariance matrix $[\sigma_{bb}]$ of $[b]$ as

$$[S_{bb}] = \begin{bmatrix} S_{b_1 b_1} & S_{b_1 b_2} & \dots & S_{b_1 b_m} \\ S_{b_2 b_1} & S_{b_2 b_2} & \dots & S_{b_2 b_m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{b_m b_1} & S_{b_m b_2} & \dots & S_{b_m b_m} \end{bmatrix}$$

(2) There are statistics $z_1, z_2, \dots, z_m$ calculated from the data of the first phase sample such that

$$w = b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]' [z]$$

and the estimator of the covariance matrix $[\sigma_{zz}]$ of $[z]$ is denoted by

$$[S_{zz}] = \begin{bmatrix} S_{z_1 z_1} & S_{z_1 z_2} & \dots & S_{z_1 z_m} \\ S_{z_2 z_1} & S_{z_2 z_2} & \dots & S_{z_2 z_m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{z_m z_1} & S_{z_m z_2} & \dots & S_{z_m z_m} \end{bmatrix}$$

(3) The vector $[b]$ is statistically independent of $[z]$. Then, it has been shown by Cunia (1965) that

(1) the variance σ_{ww} of w is given by the expression

$$\sigma_{ww} = [\beta]'[\sigma_{zz}][\beta] + [\mu_z]'[\sigma_{bb}][\mu_z] + \sum_{i=1}^m \sum_{j=1}^m \sigma_{z_i z_j} \sigma_{b_i b_j}$$

where $[\mu_z]$ is the vector of the expected values of $z_1, z_2, \dots, z_m$,

(2) an estimator of the variance σ_{ww} of w is given by

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z] - \sum_{i=1}^m \sum_{j=1}^m S_{z_i z_j} S_{b_i b_j}$$

and (3) the terms $S_{z_i z_j} S_{b_i b_j}$ are relatively small compared to the corresponding terms $b_i b_j S_{z_i z_j}$ or $z_i z_j S_{b_i b_j}$ and can ordinarily be ignored.

Consequently, the variance of w can be estimated from the sample values by the expression

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z] = \sum_{i=1}^m \sum_{j=1}^m (b_i b_j S_{z_i z_j} + z_i z_j S_{b_i b_j})$$

Note that in the above formula we have two terms and that the first can be interpreted as the error component of the first phase and the second, the error component of the second phase sample.

If one wants to ignore the error of the biomass tables, that is, assume that $[S_{bb}] = [0]$, the variance of w reduces then to the first component only; the second component vanishes. This would be the case where the biomass tables are thought to be sufficiently accurate so that their error can be ignored or the case where the error of biomass tables is not known, cannot be calculated and, thus, one has to ignore it.

There is also the case where the error of the first phase sample is equal to zero, or sufficiently close to zero, so that it may be reasonable to make $[S_{zz}] = [0]$. For example, this case may occur when one takes the inventory of a small experimental forest area and every tree can be measured for all characteristics but biomass. Then the first error component vanishes and the variance of w reduces to the second component only.

It may be interesting to see what happens if we use relative rather than absolute measures of error. If we define the "percent error" as the ratio

$$\text{percent error} = \frac{\text{standard error of } w}{w}$$

and if we divide the variance of w by w^2 , that is,

$$\frac{S_{ww}}{w^2} = \frac{[b]'[S_{zz}][b]}{w^2} + \frac{[z]'[S_{bb}][z]}{w^2}$$

we obtain the relationship

$$(\text{total percent error})^2 = (\text{percent error of first component})^2 + (\text{percent error of second component})^2$$

This shows that an estimate of the total error of w (percent error or variance) can be obtained by a simple formula, when separately calculated estimates of the error components from the two sources (the samples of the first and second phase) are available in a percent error form.

There seems to be no problem with the determination of the percent error due to the sample of the first phase. The variance of $w = [b]'[z]$ and the corresponding percent error component can be calculated by the formula

$$\sqrt{[b]'[S_{zz}][b]} / [b]'[z]$$

The problem is much more difficult, however, with the calculation of the percent error of the biomass regression function; it requires an estimate of the expected value of $[z]$ which may or may not be calculated from the second phase sample. As an example of how to approach this problem, consider the following illustrative case.

Assume that (i) the second phase sample is selected by simple random sampling, (ii) the true regression function of the tree biomass y on diameter d is of the parabolic form $\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$ (iii) all of the other basic assumptions of the weighted least squares method are satisfied and, thus, the estimator of $[\beta]$ is the usual weighted least squares estimator $[b]$ with covariance matrix estimated by $[S_{bb}]$, and (iv) an estimator of μ = average biomass per acre is required. Then, it can be shown, see Cunia (1985), that the percent error of the second phase sample can be estimated by the formula

$$\sqrt{[X]'[S_{bb}][X]} / [b]'[X]$$

or

$$\sqrt{[X]'[S_{bb}][X]} / [b]'[X]$$

where

$$[\bar{X}] = \begin{bmatrix} \sum x_1/n \\ \sum x_2/n \\ \vdots \\ \sum x_m/n \end{bmatrix} = [X]/n \text{ and } [X] = \begin{bmatrix} \sum x_1 \\ \sum x_2 \\ \vdots \\ \sum x_m \end{bmatrix}$$

with the summation sign Σ taken over the n sample trees of the second phase.

Note that, when the trees are selected for arbitrary values of $x_1, x_2, \dots, x_m$, there seems to be no way to determine the percent error from the data of the second phase sample alone.

Applications to a Simple Forest Inventory Design

Let us consider again the oversimplified example of a forest area containing trees of a given species. In the first phase, n_p sample plots of fixed area (say "a" acres) are selected by simple random sampling (with replacement) and all their trees of the given species are measured for their diameter d at breast height. If trees of a different species are encountered, they are simply ignored. There are n_h trees in the h -th plot and the diameter of the k -th tree in the h -th plot is denoted by d_{hk} , $k=1, 2, \dots, n_h$ and $h=1, 2, \dots, n_p$.

In the second phase, n_t trees of the given species are selected at random and measured for their diameter d and volume y . Let us assume that (i) the true regression function of tree volume on diameter is of the parabolic form $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$, (ii) the conditional variance of y given d is proportional to d^4 and (iii) all other basic assumptions of the classical linear regression model are satisfied. It is then known that the best linear and unbiased estimators of β_1 , β_2 and β_3 are obtained by the weighted least squares method, described in more detail by Cunia (1986a), among others.

Let us assume here that, the estimate of the regression function is

$$\hat{y} = b_1 + b_2 d + b_3 d^2 \\ = b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x]$$

where $x_1=1$, $x_2=d$ and $x_3=d^2$ and the estimate of the covariance matrix of $[b]$ is

$$[S_{bb}] = \begin{bmatrix} S_{b_1 b_1} & S_{b_1 b_2} & S_{b_1 b_3} \\ S_{b_2 b_1} & S_{b_2 b_2} & S_{b_2 b_3} \\ S_{b_3 b_1} & S_{b_3 b_2} & S_{b_3 b_3} \end{bmatrix}$$

where $S_{b_i b_j}$ is the estimate of the covariance of b_i and b_j , $i, j = 1, 2, 3$.

To calculate w , the estimate of the average volume per acre, one can use one of the following two approaches.

Approach 1 - The volume y_{hk} of the k -th tree in the h -th plot is estimated by the regression value

$$\hat{y}_{hk} = b_1 + b_2 d_{hk} + b_3 d_{hk}^2$$

The volume v_h of the plot h is estimated by the sum of the volumes $\hat{y}_{hk}$ of the n_h trees, that is,

$$\hat{v}_h = \sum \hat{y}_{hk} = \hat{y}_{h1} + \hat{y}_{h2} + \dots + \hat{y}_{hn_h} \\ = b_1 n_h + b_2 \sum d_{hk} + b_3 \sum d_{hk}^2$$

where $\sum$ denotes summation over subscript k of all the trees of the plot h . For convenience, let us define the new variables

$$s_{h1} = n_h/a = (\text{number of trees}), \text{ on a per acre basis, of plot } h$$

$$s_{h2} = \sum d_{hk}/a = (\text{sum of } d_{hk}), \text{ on a per acre basis, of plot } h$$

$$s_{h3} = \sum d_{hk}^2/a = (\text{sum of } d_{hk}^2), \text{ on a per acre basis, of plot } h$$

$$\text{and } \hat{w}_h = \hat{v}_h/a = b_1 s_{h1} + b_2 s_{h2} + b_3 s_{h3} = [b]'[s_h] \\ = \text{estimator of the volume, on a per acre basis, of plot } h$$

The average volume per acre is now estimated as

$$w = b_1 z_1 + b_2 z_2 + b_3 z_3 = [b]'[z]$$

where

$$w = \sum \hat{w}_h / n_p$$

$$z_1 = \sum s_{h1} / n_p \\ = \text{average (number of trees) per acre} = \bar{s}_1$$

$$z_2 = \sum s_{h2} / n_p \\ = \text{average (sum of diameters) per acre} = \bar{s}_2$$

$$z_3 = \sum s_{h3} / n_p \\ = \text{average (sum of squared diameters) per acre} = \bar{s}_{h3}$$

and $\sum$ denotes summation over subscript h from 1 to n_p .

Because the variances and covariance of the variables z_i and z_j are estimated by expressions of the form

$$S_{z_i z_j} = \sum (s_{hi} - \bar{s}_i)(s_{hj} - \bar{s}_j) / n_p(n_p - 1) = S_{ij} / n_p$$

the covariance matrix of the vector

$[z]' = [z_1 \ z_2 \ z_3]$ is estimated by

$$[S_{zz}] = \begin{bmatrix} S_{z_1 z_1} & S_{z_1 z_2} & S_{z_1 z_3} \\ S_{z_2 z_1} & S_{z_2 z_2} & S_{z_2 z_3} \\ S_{z_3 z_1} & S_{z_3 z_2} & S_{z_3 z_3} \end{bmatrix} = [S_{ij}] / n_p$$

We are now in a position to apply the formulae of the previous section and write that the variance of w is estimated by

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

The terms $S_{z_i z_j} S_{b_i b_j}$ of lower order of magnitude

have been left out of the formula above since their value is relatively small.

It may be of interest here to mention the fact that the common procedure for the estimation of the variance of w is by the formula

$$S_{ww} = (S_{\hat{v}\hat{v}} / n_p) / a^2$$

where

$$S_{\hat{v}\hat{v}} = (\sum \hat{v}_h^2 - (\sum \hat{v}_h)^2 / n_p) / (n_p - 1) \\ = \text{estimate of the variance of the estimates } \hat{v}_h \text{ of the individual plot volumes } v_h$$

and $\sum$ denotes summation over subscript h from 1 to n_p .

By expressing $\hat{v}_h$ as

$$\hat{v}_h = b_1 n_h + b_2 \sum d_{hk} + b_3 \sum d_k^2$$

$$= a(b_1 s_{h1} + b_2 s_{h2} + b_3 s_{h3}) = a \hat{w}_h$$

by evaluating $\sum \hat{v}_h^2$ and $(\sum \hat{v}_h)^2$ and by rearranging terms, it can be shown that

$$S_{ww} = [b]'[S_{zz}][b]$$

This means that the common procedure for the estimation of the variance of w is biased; the variance component due to the error of the biomass regression is completely ignored. Furthermore, if one wishes to ignore the error due to biomass regression, he may use either of the two formulae, that is, either

$$S_{ww} = S_{vv}/n_p a^2$$

or

$$S_{ww} = [b]'[S_{zz}][b]$$

Approach 2 - To calculate w , one can also construct first a frequency table giving the average number of trees per acre N_c for the diameter class c , where d_c is the value of the tree diameter corresponding to the class c . There are as many diameter classes as there are distinct diameter values. Then, the volume y_c corresponding to the average number N_c of trees per acre is estimated by

$$\hat{y}_c = N_c(b_1 + b_2 d_c + b_3 d_c^2)$$

and the average volume per acre (including all diameter size trees) is estimated by

$$w = \hat{y}_1 + \hat{y}_2 + \dots + \hat{y}_C + \dots + \hat{y}_M$$

$$= b_1 \sum N_c + b_2 \sum N_c d_c + b_3 \sum N_c d_c^2$$

where $\sum$ means summation over all diameter sizes. Because $\sum N_c = z_1$, $\sum N_c d_c = z_2$, and $\sum N_c d_c^2 = z_3$, one can finally write, as before with the first approach,

$$w = b_1 z_1 + b_2 z_2 + b_3 z_3 = [b]'[z]$$

Note that the above formulae do not take into account the effect of the measurement bias. If the tree volume is measured so that a bias exists in this measurement, the bias will be preserved in the estimate of w and will not be included in the estimate of the error of w . Similarly, the error due to the selection of the statistical model is not accounted for. It is certain that the parabolic regression form assumed here is not the true form and one may well use a different form and obtain completely different estimates. In all cases, there will be a discrepancy between the assumed and the true form of the regression function. The error due to this discrepancy is minimized by a judicious choice of the regression function form.

For a numerical illustration of how these formulae apply to an actual case consider the following example.

Table 1 - Species group s and diameters d (to the nearest one-tenth of an inch) of the trees of the first phase sample plot 1.

s	d	s	d	s	d	s	d	s	d	s	d
3	10.2	3	5.1	3	5.8	3	6.1	3	19.8	1	20.1
1	5.1	2	12.0	2	12.2	3	6.5	3	5.2	3	5.1
1	16.1	3	8.1	1	14.1	3	6.4	2	17.4	3	8.2
2	12.5	3	5.0	1	11.4	3	16.1	1	12.0	1	5.7
2	9.1	1	12.9	3	9.4	1	18.6	3	5.7	-	--
2	13.9	3	5.8	3	15.0	2	12.6	3	5.3	-	--

Example 1 - In the first phase, $n_p = 926$ one-fifth acre sample plots are selected at random from the New York State forest area, and all their trees are classified by species group (1 for pines, 2 for maples, and 3 for all remaining species) and measured for their diameter at breast height d to the nearest one-tenth of an inch. As an example, the species groups and diameters of the trees of sample plot 1 are listed in Table 1. In the second phase, $n_t = 353$ trees are also selected at random from the same forest area. Their species group, diameter d and total above-ground biomass y (green weight to the nearest pound) are listed by Cunia (1986a).

A statistical analysis of the data of the second phase sample showed that, for all species groups combined,

(1) a parabolic function of the form

$$\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = [b]'[x]$$

is a satisfactory expression for the regression function of the tree biomass y on diameter d , where $x_1 = 1$, $x_2 = d$, and $x_3 = d^2$, and

(2) the conditional variance of y given d is approximately proportional to the fourth power of d , that is, for unknown σ^2

$$\sigma_{yy/x} = \sigma^2 d^4$$

Using these assumptions (in addition to the other assumptions of the linear least squares regression method) and the sample data of the 353 trees, Cunia (1986a) shows that, for the three species combined,

$$[b] = \begin{bmatrix} 5.1818118 \\ -25.653078 \\ 12.988357 \end{bmatrix} \quad \text{and}$$

$$[S_{bb}] = \begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

To calculate an estimate of the average biomass per acre for all species combined, when the same biomass regression function is used for all species, we shall apply first the usual procedure consisting of the following steps

(1) Each individual tree biomass $\hat{y}_{hk}$, where $h = 1, 2, \dots, 926$ is the plot number and $k = 1,$

2, ..., n_h is the tree number within the given plot h , is estimated by the regression function as

$$\hat{y}_{hk} = b_1 + b_2 d_{hk} + b_3 d_{hk}^2$$

For example, to estimate the biomass of the trees of plot 1 of Table 1, we calculate successively

$$y_{11} = 5.181812 - (25.653078)(10.2) + (12.988357)(10.2)^2 = 1094.8291$$

= estimate of the biomass of the first tree of plot 1

$$y_{12} = 5.181812 - (25.653078)(5.1) + (12.988357)(5.1)^2 = 212.17827$$

= estimate of the biomass of the second tree of plot 1

.

.

$$y_{1,34} = 5.181812 - (25.653078)(5.7) + (12.988357)(5.7)^2 = 280.95098$$

= estimate of the biomass of the last tree of plot 1

(2) Each individual plot biomass $\hat{v}_h$, $h = 1, 2, \dots, 926$ is calculated by summing up the individual tree biomass within the plot. For example, the biomass of the first sample plot is estimated by the value

$$\hat{v}_1 = 1094.8291 + 212.17827 + \dots + 280.95098 = 48845.674$$

(3) By calculating first

$$\hat{\Sigma v}_h = 22312916 \text{ and } \hat{\Sigma v}_h^2 = 963042020000$$

where Σ means summation over $h = 1, 2, \dots, 926$, we calculate successively

$$\bar{v} = \hat{\Sigma v}_h / n_p = 22312916 / 926 = 24096.022$$

= estimate of the average biomass per plot, say μ_v ,

$$S_{vv} = \Sigma(\hat{v}_h - \bar{v})^2 / (n_p - 1) = 459880550$$

= estimate of the variance of the estimated plot biomass v_h

$$S_{\bar{v}\bar{v}} = S_{vv} / n_p = 496631.26$$

= estimate of the variance of $\bar{v}$

It is customary to report these estimates on a "per acre" basis. As the plot size is one-fifth of an acre, we can write

$$w = \bar{v} / (1/5) = 120480.11$$

= estimate of the average biomass per acre

and

$$S_{ww}^{(1)} = S_{\bar{v}\bar{v}} / (1/5)^2 = 12415782$$

= estimate of the variance of w

It is also customary to calculate the 95 percent confidence limits of μ , the true average biomass per acre. Using, for convenience, a t -value of 2, we have calculated the 95 percent interval as equal to

$$w \pm 2\sqrt{S_{ww}^{(1)}} = 120480 \pm 7047$$

This statement implies that the true mean μ lies, with a .95 confidence, somewhere between a lower limit of $120480 - 7047 = 113433$ and an upper limit of $120480 + 7047 = 127527$ pounds.

However, it can be shown that the value $S_{ww}^{(1)}$ as calculated above contains only the error of the first phase sample; the error of the biomass regression function calculated from the tree data of the second phase sample is being ignored. We shall now calculate the same value w by the procedure outlined in this paper and show how we can include the error components from the samples of both phases, when estimating the error of w . Note that the superscript (1) of $S_{ww}^{(1)}$ refers to the error component of the first phase; we shall use superscript (2) to denote the corresponding error component of the second phase.

We start with the calculation of the individual plot variables defined as

$$s_{h1} = (\text{number of trees}) \text{ per acre of plot } h$$

$$s_{h2} = (\text{sum of tree diameters}) \text{ per acre of plot } h, \text{ and}$$

$$s_{h3} = (\text{sum of squared tree diameters}) \text{ per acre of plot } h$$

For example, for the sample data of plot 1 of Table 1 we have

$$s_{11} = n_1 / (1/5) = (5)(34) = 170$$

$$s_{12} = (\Sigma d) / (1/5) = (5)(354.5) = 1772.5$$

$$s_{13} = (\Sigma d^2) / (1/5) = (5)(4447.33) = 22236.65$$

The biomass of the plot h expressed on a "per acre" basis is estimated by

$$\hat{w}_h = [b]'[s_h] = b_1 s_{h1} + b_2 s_{h2} + b_3 s_{h3}$$

For plot 1 we have

$$\hat{w}_1 = (5.1818118)(170) - (25.653078)(1772.5) + (12.988357)(22236.65) = 244228.37$$

As the reader can verify, the value $\hat{w}_1$ found here is five times larger than the value $\hat{v}_1$ found above. There is no need, however, to estimate the biomass of each individual sample plot.

Using standard procedures, we calculate the sample means, variances and covariances of the variables s_{hi} , $i = 1, 2, 3$, for the sample plot data of the first phase. Then, for Σ denoting summation over $h = 1, 2, \dots, 926$, we have

$$\bar{s}_1 = \Sigma s_{h1} / n_p = 114540 / 926 = 123.69330$$

$$\bar{s}_2 = \Sigma s_{h2}/n_p = 1009986/926 = 1090.6982$$

$$\bar{s}_3 = \Sigma s_{h3}/n_p = 1053694/926 = 11380.879$$

$$S_{11} = \Sigma (s_{h1} - \bar{s}_1)^2 / (n_p - 1) = 9610968.9/925 = 10390.237$$

$$S_{12} = \Sigma (s_{h1} - \bar{s}_1)(s_{h2} - \bar{s}_2) / (n_p - 1) = 83216060/925 = 89963.308$$

$$S_{13} = \Sigma (s_{h1} - \bar{s}_1)(s_{h3} - \bar{s}_3) / (n_p - 1) = 799124260/925 = 863918.12$$

$$S_{22} = \Sigma (s_{h2} - \bar{s}_2)^2 / (n_p - 1) = 767822580/925 = 830078.46$$

$$S_{23} = \Sigma (s_{h2} - \bar{s}_2)(s_{h3} - \bar{s}_3) / (n_p - 1) = 7975575200/925 = 8622243.4$$

$$S_{33} = \Sigma (s_{h3} - \bar{s}_3)^2 / (n_p - 1) = 91041895000/925 = 9842367$$

As $z_i = \bar{s}_i$ and $S_{z_i z_j} = S_{ij}/n_p$ for $i, j = 1, 2, 3$, we can write immediately the vector $[z]$ and the matrix $[S_{zz}]$, the estimate of the covariance matrix of $[z]$ as

$$[z] = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 123.69330 \\ 1090.6982 \\ 11380.879 \end{bmatrix} = \text{estimate of the average vector } [\mu_z]$$

and

$$[S_{zz}] = \begin{bmatrix} 11.220558 & 97.152600 & 932.95694 \\ 79.152600 & 896.41303 & 9311.2780 \\ 932.95694 & 9311.2780 & 106289.06 \end{bmatrix}$$

= estimate of the covariance matrix $[\sigma_{zz}]$ of $[z]$

We are now ready to calculate

$$w = [b]'[z] = (5.1818118)(123.69330) - (25.653078)(1090.6982) + (12.988357)(11380.879)$$

= 120480.1084 = estimate of the average biomass per acre

$$S_{ww}^{(1)} = [b]'[S_{zz}][b] = 12415782$$

= first variance component (due to first phase sample)

$$S_{ww}^{(2)} = [z]'[S_{bb}][z] = 6643784.4$$

= second variance component (due to second phase sample)

$$S_{ww}^{(3)} = \Sigma \Sigma s_{bij} s_{zizj} = 11292.170$$

= component of the error that is being ignored

As the reader can verify (i) w and $S_{ww}^{(1)}$ are

the same values as those obtained before, (ii) the values of the variance terms (of the third component) that are being ignored are small with respect to the first and the second component, and (iii) an estimate of the variance of w is

$$S_{ww} = S_{ww}^{(1)} + S_{ww}^{(2)} = 19059566$$

It may be interesting to show the additivity of the relative errors due to the first and second phase samples. If

$$100\sqrt{[b]'[S_{zz}][b]}/w = 100\sqrt{12415782/120480.89} = 2.9246161$$

= percent error due to the sample plots of the first phase

$$100\sqrt{[z]'[S_{bb}][z]}/w = 100\sqrt{6643784.4/120480.89} = 2.1393882$$

= percent error due to the sample trees of the second phase

and

$$100\sqrt{S_{ww}}/w = 100\sqrt{19059566/120480.89} = 3.623584$$

= total percent error

then, one can verify that

$$\begin{aligned} & (\text{percent error of sample plots})^2 \\ & + (\text{percent error of sample trees})^2 \\ & = 8.5533795 + 4.576982 = 13.13036 \\ & = (3.623584)^2 \\ & = (\text{total percent error})^2 \end{aligned}$$

Using a t -value of 2, we can calculate the 95 percent confidence limits as

$$w \pm 2\sqrt{S_{ww}} = 120480 \pm 8731$$

that is, a lower limit of $120480 - 8731 = 111749$ and an upper limit of $120480 + 8731 = 129211$

Note that ignoring the error in the biomass regression function leads to an underestimation of the variance by

$$(100) \left(\frac{19059566 - 12415782}{19059566} \right) = \frac{664378400}{19059566} = 34.86 \text{ percent}$$

In terms of standard error, the underestimation is equal to

$$(100) \left(\frac{\sqrt{19059566} - \sqrt{12415782}}{\sqrt{19059566}} \right) = 19.29 \text{ percent}$$

Extension to the Case of More than One Species

The extension to more than one, say $p > 1$ species is straightforward. We start by defining the "giant size" vector $[B]$ containing the p individual species vectors of regression coefficients $[b^i]$, $i = 1, 2, \dots, p$

$$[B]' = \begin{bmatrix} [b^1]' & [b^2]' & \dots & [b^p]' \end{bmatrix}$$

The covariance matrix of $[B]$ is estimated by the "giant size" matrix

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [S_{bb}^{12}] & \dots & [S_{bb}^{1p}] \\ [S_{bb}^{12}]' & [S_{bb}^{22}] & \dots & [S_{bb}^{2p}] \\ \vdots & \vdots & \ddots & \vdots \\ [S_{bb}^{1p}]' & [S_{bb}^{2p}]' & \dots & [S_{bb}^{pp}] \end{bmatrix}$$

where $[S_{bb}^{ij}]$ is the estimate of the covariance matrix of $[b^i]$ and $[b^j]$, $i, j = 1, 2, \dots, p$. When $[b^i]$ and $[b^j]$ are statistically independent, $[S_{bb}^{ij}] = [0]$. This may be the case when biomass regressions are estimated separately by species from statistically independent samples. When the samples are not independent, or the regression functions are not calculated independently of each other, the problem of calculating $[S_{bb}^{ij}]$ may become difficult, if at all possible, to solve.

We define now the mp variables $s_{11}, s_{12}, \dots, s_{pm}$ for each sample plot h. The first m belong to the first, the next m to the second, and the last m to the p-th species. The m variables of each species denote the sums of values $x_1, x_2, \dots, x_m$ of plot h expressed on a per acre basis and they are calculated as shown above for a single species. In the same way, the averages per acre values $z_{11}, z_{12}, \dots, z_{pm}$ are calculated as well as their covariance matrices $[S_{zz}^{ij}]$. This yields the vector $[Z]$ of estimates and covariance matrix $[S_{ZZ}]$, that is

$$[Z]' = \begin{bmatrix} [z^1]' & [z^2]' & \dots & [z^p]' \end{bmatrix}$$

and

$$[S_{ZZ}] = \begin{bmatrix} [S_{zz}^{11}] & [S_{zz}^{12}] & \dots & [S_{zz}^{1p}] \\ [S_{zz}^{12}]' & [S_{zz}^{22}] & \dots & [S_{zz}^{2p}] \\ \vdots & \vdots & \ddots & \vdots \\ [S_{zz}^{1p}]' & [S_{zz}^{2p}]' & \dots & [S_{zz}^{pp}] \end{bmatrix}$$

Because each plot has mp variables s, the covariances $[S_{zz}^{ij}]$ are generally different from zero.

Consequently, one can use the formulae of the previous section and obtain

$$w = [B]'[Z]$$

= estimate of the average volume per acre (total of the p species)

and

$$S_{ww} = [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z]$$

= estimate of the variance of w

Note that one can calculate, if he so wishes, the average value per acre for a given species. For example, for the first species one

can define

$$w_1 = [b^1]'[z^1] = \text{estimate of the average volume per acre for the first species,}$$

where

$$[b^1]' = [b_{11} \quad b_{12} \quad \dots \quad b_{1m}] \text{ and } [z^1]' = [z_{11} \quad z_{12} \quad \dots \quad z_{1m}]$$

The variance of w_1 is estimated as usual, by the formula

$$S_{w_1 w_1} = [b^1]'[S_{zz}^{11}][b^1] + [z^1]'[S_{bb}^{11}][z^1]$$

where $[S_{zz}^{11}]$ is the covariance matrix of $[z^1]$.

Furthermore, the covariance of w_i and w_j can be estimated by the formula

$$S_{w_i w_j} = [b^i]'[S_{zz}^{ij}][b^j] + [z^i]'[S_{bb}^{ij}][z^j]$$

To illustrate how the above formulae apply to an actual case, consider the following numerical example.

Example 2 - Using the sample data of Example 1, let us calculate the estimates of (i) the average biomass per acre for each species group separately, and (ii) the average biomass per acre of all species combined when three, statistically independent biomass regression functions are calculated from the data of the first phase sample, one for each species group. In this last case we shall also compare the present estimates with those obtained in Example 1 where a single biomass regression function is used for all three species groups.

Applying to each species group separately the weighted least squares method, Cunia (1986b) gives the following statistics from the second phase sample, where the superscript refers to the species group number

$$[b^1] = \begin{bmatrix} 295.60183 \\ -107.06967 \\ 16.882552 \end{bmatrix}, [b^2] = \begin{bmatrix} -256.70604 \\ 40.050701 \\ 9.1695394 \end{bmatrix}, [b^3] = \begin{bmatrix} 18.800242 \\ -20.693393 \\ 13.156786 \end{bmatrix}$$

$$[S_{bb}^{11}] = \begin{bmatrix} 13256.622 & -3494.4363 & 211.95740 \\ -3494.4363 & 943.83938 & -58.826156 \\ 211.95740 & -58.826156 & 3.8046237 \end{bmatrix}$$

$$[S_{bb}^{22}] = \begin{bmatrix} 25911.067 & -6691.9889 & 393.93612 \\ -6691.9889 & 1777.8435 & -108.29003 \\ 393.93612 & -108.29003 & 6.9277172 \end{bmatrix}$$

$$[S_{bb}^{33}] = \begin{bmatrix} 25197.692 & -6243.4168 & 347.66044 \\ -6243.4168 & 1587.0539 & -91.114812 \\ 347.66044 & -91.114812 & 5.4758847 \end{bmatrix}$$

For convenience, we shall arrange the nine regression coefficients in a giant-size vector [B] defined as

$$[B] = \begin{bmatrix} [b^1] \\ [b^2] \\ [b^3] \end{bmatrix}$$

and the covariance matrix of [B] is the giant-size matrix

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [0] & [0] \\ [0] & [S_{bb}^{22}] & [0] \\ [0] & [0] & [S_{bb}^{33}] \end{bmatrix}$$

where [0] denotes a 3 by 3 zero matrix.

The corresponding statistics of the first phase sample are similarly calculated. For each sample plot there are now nine variables, three for each species group. If the superscript refers again to the species group number, and if, as an example, we use the data of the plot 1 given in Table 1, we can write

$$s_{11}^1 = (9)/(1/5) = 45 = \text{number of trees per acre of species group 1}$$

$$s_{12}^1 = (116)/(1/5) = 580 = \text{sum of diameters per acre of species group 1}$$

.

$$s_{33}^3 = (1553.84)/(1/5) = 7769.20 = \text{sum of squared tree diameters per acre of species group 3}$$

Note that the first subscript 1, 2 or 3 refers to the species group 1, 2 or 3, while the second subscript 1, 2 and 3 refers to the variable s_1 , s_2 and s_3 respectively.

The averages of the nine plot variables are the elements of the three species group vectors

$$[z^1] = \begin{bmatrix} 28.439525 \\ 245.73704 \\ 2592.1076 \end{bmatrix} = \begin{bmatrix} \text{average number of trees per acre of species group 1} \\ \text{average sum of diameters per acre of species group 1} \\ \text{average sum of (diameters)}^2 \text{ per acre of species group 1} \end{bmatrix}$$

and similarly defined

$$[z^2] = \begin{bmatrix} 39.994600 \\ 352.64039 \\ 3665.4591 \end{bmatrix} \quad \text{for species group 2}$$

and

$$[z^3] = \begin{bmatrix} 55.259179 \\ 492.32073 \\ 5133.3124 \end{bmatrix} \quad \text{for species group 3}$$

For convenience, the nine averages z are arranged in the giant-size vector [Z] defined as

$$[Z] = \begin{bmatrix} [z^1] \\ [z^2] \\ [z^3] \end{bmatrix}$$

To calculate the covariance matrix of the giant-size vector [Z], we shall use the procedure applied in Example 1 to a vector [z] of size 3. If $[S_{zz}^{ij}]$ denotes the covariance matrix of vector $[z^i]$ with vector $[z^j]$, that is, if

$$[S_{zz}^{ij}] = \begin{bmatrix} s_{z_1 z_1}^{ij} & s_{z_1 z_2}^{ij} & s_{z_1 z_3}^{ij} \\ s_{z_2 z_1}^{ij} & s_{z_2 z_2}^{ij} & s_{z_2 z_3}^{ij} \\ s_{z_3 z_1}^{ij} & s_{z_3 z_2}^{ij} & s_{z_3 z_3}^{ij} \end{bmatrix}$$

we have the sample value of the estimate of the giant size covariance matrix of [Z] equal to

$$[S_{ZZ}] = \begin{bmatrix} [S_{zz}^{11}] & [S_{zz}^{12}] & [S_{zz}^{13}] \\ [S_{zz}^{21}] & [S_{zz}^{22}] & [S_{zz}^{23}] \\ [S_{zz}^{31}] & [S_{zz}^{32}] & [S_{zz}^{33}] \end{bmatrix}$$

where

$$[S_{zz}^{11}] = \begin{bmatrix} 3.3631663 & 30.013346 & 299.99303 \\ 30.013376 & 278.99377 & 2924.2532 \\ 299.99303 & 2924.2532 & 32478.880 \end{bmatrix}$$

= estimate of the covariance matrix of $[z^1]$

$$[S_{zz}^{12}] = \begin{bmatrix} -.13931796 & -1.1649666 & -12.653270 \\ -.85774773 & -6.1741338 & -58.815740 \\ -10.140402 & -65.157745 & -488.21114 \end{bmatrix}$$

= estimate of the covariance matrix of $[z^1]$ with $[z^2]$

$$[S_{zz}^{13}] = \begin{bmatrix} -.028807992 & -.17263912 & -2.4404317 \\ .26017525 & 3.8362467 & 43.326113 \\ .78247233 & 28.612932 & 406.69373 \end{bmatrix}$$

= estimate of the covariance matrix of $[z^1]$ with $[z^3]$

$$[S_{zz}^{22}] = \begin{bmatrix} 3.3088844 & 27.803629 & 263.44704 \\ 27.803629 & 249.16870 & 2570.5596 \\ 263.44704 & 2570.5596 & 29632.556 \end{bmatrix}$$

= estimate of the covariance matrix of $[z^2]$

$$[S_{zz}^{23}] = \begin{bmatrix} .57702562 & 6.0188741 & 69.996561 \\ 4.0090757 & 46.423683 & 590.72494 \\ 27.971175 & 388.40702 & 5613.8833 \end{bmatrix}$$

= estimate of the covariance matrix of $[z^2]$ with $[z^3]$

$$[S_{zz}^{33}] = \begin{bmatrix} 3.7307078 & 31.242824 & 296.00076 \\ 31.242824 & 280.07896 & 2889.3678 \\ 296.00076 & 2889.3678 & 33112.894 \end{bmatrix}$$

= estimate of the covariance matrix of $[z^3]$

$[S_{zz}^{21}] = [S_{zz}^{12}]'$ = estimate of the covariance matrix of $[z^2]$ with $[z^1]$
 $[S_{zz}^{31}] = [S_{zz}^{13}]'$ = estimate of the covariance matrix of $[z^3]$ with $[z^1]$
 and

$[S_{zz}^{32}] = [S_{zz}^{23}]'$ = estimate of the covariance matrix of $[z^3]$ with $[z^2]$

We have now all the elements needed for the calculation of the average biomass per acre for each species group separately and their sum. Using the usual formulae, we find the following statistics:

$$w_1 = [b^1]'[z^1] = 25857.183 = \text{average biomass per acre of species group 1}$$

$$w_2 = [b^2]'[z^2] = 37375.515 = \text{average biomass per acre of species group 2}$$

$$w_3 = [b^3]'[z^3] = 58388.993 = \text{average biomass per acre of species group 3}$$

$$w = [B]'[Z] = 121621.69 = \text{average biomass per acre of all species combined}$$

$$= 25857.183 + 37375.515 + 58388.993$$

$$= w_1 + w_2 + w_3$$

The variances of w_1 , w_2 , w_3 and $w = w_1 + w_2 + w_3$ are also calculated by the usual formulae as

$$S_{ww}^{11} = [b^1]'[S_{zz}^{11}][b^1] + [z^1]'[S_{bb}^{11}][z^1] \\ = 3271962.1 + 746466.01 = 4018428.1$$

$$S_{ww}^{22} = [b^2]'[S_{zz}^{22}][b^2] + [z^2]'[S_{bb}^{22}][z^2] \\ = 3185350.8 + 2338326.6 = 5523677.4$$

$$S_{ww}^{33} = [b^3]'[S_{zz}^{33}][b^3] + [z^3]'[S_{bb}^{33}][z^3] \\ = 4401941.3 + 2899185.7 = 7301127.0$$

and

$$S_{ww} = [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z] \\ = 12145981 + 5983978.3 = 18129960$$

As the reader can verify,

$$S_{ww} \neq S_{ww}^{11} + S_{ww}^{22} + S_{ww}^{33} = 16843232$$

This is because w_1 , w_2 , and w_3 are not statistically independent. The right relationship contains also the covariances of w_1 , w_2 , and w_3 , more specifically

$$S_{ww}^{12} = [b^1]'[S_{zz}^{12}][b^2] + [z^1]'[S_{bb}^{12}][z^2] \\ = -52560.39 + 0 \quad (\text{since } [S_{bb}^{12}] = [0])$$

$$S_{ww}^{13} = [b^1]'[S_{zz}^{13}][b^3] + [z^1]'[S_{bb}^{13}][z^3] \\ = 18934.588 + 0 \quad (\text{since } [S_{bb}^{13}] = [0])$$

and

$$S_{ww}^{23} = [b^2]'[S_{zz}^{23}][b^3] + [z^2]'[S_{bb}^{23}][z^3] \\ = 876989.48 + 0 \quad (\text{since } [S_{bb}^{23}] = [0])$$

The reader can now verify that

$$S_{ww} = S_{ww}^{11} + S_{ww}^{22} + S_{ww}^{33} + 2(S_{ww}^{12} + S_{ww}^{13} + S_{ww}^{23}) \\ = 18129960$$

Using a t-value of 2, the 95 percent confidence limits are now calculated as

$$w \pm 2\sqrt{S_{ww}} = 121622 \pm 8516$$

It may be interesting to find out what the error will be when the error of the biomass regression functions is being ignored. Then,

$$S_{ww} = [B]'[S_{ZZ}][B] = 12145981$$

This means that the variance is underestimated by about

$$(100)(5983978.3/18129960) = 33.01 \text{ percent}$$

Let us now compare the results of Examples 1 and 2. There is a difference between the two estimates of the average biomass per acre of

$$(121621.69 - 120480.89) = 1140.80$$

That is, about .9 percent. Because the biomass regression functions are constructed by species, we would expect the estimate w of Example 3 to be more precise than that of Example 1. The increase in precision is not large, however. It is estimated as the difference between the two variances, that is,

$$(19059566 - 18129960) = 929606$$

or about 4.8 percent. This is to say that using biomass regressions by species improves the estimate of the average biomass per acre by less than 5 percent, a relatively small improvement.

It should be noted here, however, that the classification of the species of our artificial population into three species groups was made rather arbitrarily. We simply wanted to have a set of three sufficiently large subsamples of trees to allow us the calculation of three separate regressions. We did not group the species according to their similarities. As a result, the conditional variance of the tree biomass within species groups was found to be only slightly smaller, on the average than the conditional variance for all species combined.

An Illustrating Example Using Sample Points

We shall consider now a sampling design which, with the exception of the type of sampling unit of the first phase, is identically defined to the design of the previous sections. Instead of using fixed area sample plots, it is common to use relascope sample points, where sample trees are selected with probability proportional to their basal area (or, what is the same thing, to

their squared diameter) and the number of trees counted at a given point multiplied by the basal area expansion factor (BAF), say c , represents a measure of the basal area in square feet per acre. With a little change in the procedure for calculating the variables s_{h1} , s_{h2} , and s_{h3} of the sample point h , $h = 1, 2, \dots, n_p$, the approach described in the previous section for the calculation of w and its error can be applied to the present sampling design as well.

More specifically, let us define the point variables

$$\begin{aligned}s_{h1} &= \Sigma(1/a_{hk}) \\ s_{h2} &= \Sigma(d_{hk}/a_{hk}), \text{ and} \\ s_{h3} &= \Sigma(d_{hk}^2/a_{hk})\end{aligned}$$

where Σ is taken over all the trees counted by relascope at the sample point h , and a_{hk} is the value by which the measurements of the hk -th sample tree are divided to bring these measurements to a per acre basis. Basically a_{hk} is the area of the imaginary plot by which all trees of diameter d_{hk} are being sampled. Its value can be found by going back to the theory of point sampling. Or its value can be determined by the following indirect method.

If c represents the known BAF by which the number of trees counted at the point is multiplied to yield the basal area per acre at that point and because the tree basal area is equal to $\pi d^2/4$, then we have to solve for a_{hk} the relationship

$$(\pi d^2/4)/a_{hk} = c$$

This yields the value

$$a_{hk} = \pi d^2/4c$$

In this formula one must express d in an appropriate unit of measurement so that the unit of a_{hk} is the acre. Recall that c represents the number of square feet of basal area per acre corresponding to a given tree. In North America, the diameter is measured in inches and the basal area in square feet. Then we must express the value of a_{hk} as

$$\begin{aligned}a_{hk} &= \pi (d_{hk}/12)^2 \text{ square feet}/4c(\text{square feet/acre}) \\ &= (\pi d_{hk}^2/576c) \text{ acres}\end{aligned}$$

The reader can verify the formula of a_{hk} by assuming that the tree measurement of interest is the tree basal area. Then, for a tree of diameter d (expressed in the appropriate unit) we have

$$(\text{tree basal area})/(\pi d^2/4c) = (\pi d^2/4)/\pi d^2/4c = c$$

In Europe, however, one uses the metric system. Let us assume first that d is expressed in meters. Then

$$\begin{aligned}a_{hk} &= \pi (d_{hk} \text{ meters})^2/(4c(\text{square meters/hectare})) \\ &= (\pi d_{hk}^2/4c) \text{ hectares}\end{aligned}$$

But if d_{hk} is measured in centimeters, then we must write

$$\begin{aligned}a_{hk} &= \frac{\pi (d_{hk} \text{ centimeters}/(100 \text{ centimeters/meter}))^2}{4c(\text{square meters/hectare})} \\ &= \frac{\pi d_{hk}^2 \text{ square meters}/10000}{4c \text{ square meters/hectare}} \\ &= (\pi d_{hk}^2/40000c) \text{ hectares}\end{aligned}$$

From here on, everything is the same as for the case with fixed area plot of the previous section, that is

$$\begin{aligned}\hat{w}_h &= b_1 s_{h1} + b_2 s_{h2} + b_3 s_{h3} = [b]'[s_h] \\ &= \text{volume on a per acre basis at the sample point } h\end{aligned}$$

$$w = b_1 z_1 + b_2 z_2 + b_3 z_3 = [b]'[z]$$

where w , z_1 , z_2 , and z_3 have the same definitions and where the covariance terms $S_{z_i z_j}$ are calculated by the same formulae. This yields the variance of w , the same as before,

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z] + \text{terms whose values are usually ignored.}$$

Let us illustrate now the application of these formulae to a numerical case.

Example 3 - In the first phase 4313 clusters of ten Bitterlich point samples (with each point sample taken with a factor $c = 37.5$ square feet/acre) are selected at random from the New York State forest area, and all the trees counted by the relascope are measured, among other things, for their diameters d at breast height (nearest one-tenth of an inch). In the second phase $n_t = 353$ trees are also selected at random and their diameter d and total above ground biomass y (green weight to the nearest pound) are both measured. Let Table 2 list the tree diameters of the first cluster and assume that the trees of the second phase sample are those used in Example 1 for which the biomass regression function of the parabolic form has already been

Table 2 - Diameters d (to the nearest one-tenth of an inch) of the trees counted at the first phase cluster number 1 of ten Bitterlich sample points.

Point	d	Point	d	Point	d
1	--	5	10.7	6	11.2
2	16.3	5	7.8	7	14.8
3	5.0	5	5.2	7	10.0
3	6.8	5	10.0	8	--
3	10.3	5	9.5	9	7.4
4	10.3	6	6.2	9	6.5
4	9.6	6	10.8	9	10.3
4	9.9	6	7.8	10	6.8
5	8.7	6	11.8	10	8.3
5	6.5	6	11.7	-	--

calculated. For the numerical values of the statistics $[b]$ and $[S_{bb}]$ that are needed here, the reader is referred to Example 1. Let us then calculate an estimate w of the average biomass per acre μ and its variance S_{ww} that includes the error from both first and second phase samples.

We start by calculating the cluster variables

$$\begin{aligned} s_{h1} &= \Sigma (1/10a_{hk}) = (\text{number of trees}) \text{ per acre at cluster } h \\ s_{h2} &= \Sigma (d_{hk}/10a_{hk}) = (\text{sum of tree diameters}) \text{ per acre at cluster } h \\ \text{and} \\ s_{h3} &= \Sigma (d_{hk}^2/10a_{hk}) = (\text{sum of squared diameters}) \text{ per acre at cluster } h, \end{aligned}$$

where Σ is taken over the trees k , $k = 1, 2, \dots$, n_h counted at the cluster h of size n_h , divisor a_{hk} for a diameter d_{hk} is equal to $(\pi d_{hk}^2/576c)$, c is the BAF of 37.5 and 10 is the number of point samples in a cluster. Note that the three variables s_{h1} , s_{h2} , and s_{h3} can also be written as

$$\begin{aligned} s_{h1} &= \Sigma (576c/10\pi d_{hk}^2) = 687.54935 \Sigma (1/d_{hk}^2) \\ s_{h2} &= \Sigma (576c/10\pi d_{hk}) = 687.54935 \Sigma (1/d_{hk}) \text{ and} \\ s_{h3} &= \Sigma (576c/10\pi) = 687.54935 n_h \end{aligned}$$

As an illustration, the reader can verify that, for cluster number 1 (of Table 2) we obtain the values

$$\begin{aligned} s_{11} &= 687.54935(1/(16.3)^2 + 1/(5.0)^2 + \dots + 1/(8.3)^2) = 275.67758 \\ s_{12} &= 687.54935(1/16.3 + 1/5.0 + \dots + 1/8.3) = 2171.6585 \\ s_{13} &= (687.54935)(1 + 1 + \dots + 1) = 18563.833 \end{aligned}$$

These represent the measured values at cluster 1 of the number of trees per acre, sum of tree diameters per acre and sum of squared tree diameters per acre respectively.

To calculate the estimate w of the average biomass per acre μ and the estimate S_{ww} of its variance, we shall apply from here on, the procedure already used in Example 1. More specifically, we proceed as follows.

(1) We calculate first the sums, sums of squares and sums of cross products of the variables s_{h1} , s_{h2} , and s_{h3} as

$$\begin{aligned} \Sigma s_{h1} &= 358160.27, \Sigma s_{h2} = 3077217.2, \\ \Sigma s_{h3} &= 30776084, \\ \Sigma s_{h1}^2 &= 71609667, \Sigma s_{h1}s_{h2} = 601337690, \\ \Sigma s_{h1}s_{h3} &= 5822135900, \\ \Sigma s_{h2}^2 &= 5164432500, \Sigma s_{h2}s_{h3} = 51276564000, \\ \Sigma s_{h3}^2 &= 524756860000 \end{aligned}$$

where Σ is taken over all 4313 clusters h of the first phase sample. Using these values it is then

easy to calculate

$$\begin{aligned} \bar{s}_1 &= (358160.27)/(4313) = 83.042029 \\ &= \text{estimate of the average number of trees per acre} \\ S_{11} &= (71609667 - (358160.27)^2/4313)/(4312) \\ &= 9709.4878 \\ &= \text{estimate of the variance of the cluster number of trees per acre} \\ S_{12} &= (601337690 - (358160.27)(3077217.2)/4313)/(4312) \\ &= 80194.650 = \text{estimate of the covariance of the two cluster variables } s_{h1} \text{ (the number of trees per acre) and } s_{h2} \text{ (the sum of tree diameters per acre)} \end{aligned}$$

and similarly

$$\begin{aligned} \bar{s}_2 &= 713.47489 = \text{estimate of the average sum of tree diameters per acre} \\ \bar{s}_3 &= 7135.6560 = \text{estimate of the average sum of squared tree diameters per acre} \\ S_{22} &= 688523.95 = \text{estimate of the variance of } s_{h2} \text{ (sum of diameters per acre)} \\ S_{13} &= 757520.27 = \text{estimate of the covariance of } s_{h1} \text{ and } s_{h3} \\ S_{33} &= 70767465 = \text{estimate of the variance of } s_{h3} \text{ (sum of squared diameters per acre)} \\ \text{and} \\ S_{23} &= 6799304.4 = \text{estimate of the covariance of } s_{h2} \text{ and } s_{h3} \end{aligned}$$

Dividing the variance and covariance terms by the total number 4313 of clusters in the first phase sample, and arranging all the values in a vector and matrix form, we obtain

$$[z] = \begin{bmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \bar{s}_3 \end{bmatrix} = \begin{bmatrix} 83.042029 \\ 713.47489 \\ 7135.6560 \end{bmatrix}$$

and

$$[S_{zz}] = \begin{bmatrix} 2.2512144 & 18.593705 & 175.63651 \\ 18.593705 & 159.63922 & 1576.4675 \\ 175.63651 & 1576.4675 & 16407.945 \end{bmatrix}$$

(2) We already know, from the calculations of Example 1, the values $[b]$ and $[S_{bb}]$. Then, we can further calculate

$$\begin{aligned} w &= [b]'[z] = 74807.927 \text{ pounds per acre } \mu \\ &= \text{estimate of the average biomass per acre } \mu \\ S_{ww}^{(1)} &= [b]'[S_{zz}][b] = 1841261.6 \\ &= \text{estimate of the error component due to the first phase sample} \end{aligned}$$

$$S_{ww}^{(2)} = [z]'[S_{bb}][z] = 2145377.6$$

= estimate of the error component due to the second phase sample

$$S_{ww}^{(3)} = \sum_i \sum_j S_{b_i b_j} S_{z_i z_j} = 1119.6811$$

= estimate of the third error component which is being ignored

$$S_{ww} = S_{ww}^{(1)} + S_{ww}^{(2)} = 1841261.6 + 2145377.6 = 3986639.1$$

= estimate of the variance of w

$$2\sqrt{S_{ww}} = (2)(1996.656985) = 3993.3140$$

= half-width of the 95 percent confidence interval of μ

Consequently, the point and 95 percent confidence interval of μ , the average biomass per acre are equal to

$$w \pm 2\sqrt{S_{ww}} = (74808 \pm 3993) \text{ pounds}$$

It may be interesting to note in this example that with a phase 1 sample of 4313 clusters and a phase 2 sample of 353 trees, more than 50 percent of the error (expressed as variance) is associated with the biomass regressions. If the error of the biomass regression is ignored, the variance of w is underestimated by

$$(2145377.6)(100)/(3986639.1) = 53.8 \text{ percent}$$

If the underestimation is calculated in terms of standard error (or confidence interval) then we obtain

$$(1996.6570 - \sqrt{1841261.6})(100)/(1996.6570) = 32.0 \text{ percent}$$

It may also be interesting to note that the same biomass regression may yield different size error components in both, absolute and relative terms. In Example 1 the absolute value of the error component was estimated as equal to 6643784. This is over three times as large as the corresponding value of 2145377.6 of the present example. The relative value as found in Example 1 is

$$100\sqrt{6643784}/120480.89 = 2.14 \text{ percent}$$

while in the present example the relative value is

$$100\sqrt{2145377.6}/74807.927 = 1.96 \text{ percent}$$

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

Cunia, T. Some theory on the reliability of volume estimates in a forest inventory sample. *Forest Science*, 11:115-128; 1965.

Cunia, T. On the error of biomass estimates in forest inventories: Part 1: Its major components. *Faculty of Forestry Miscellaneous Publication Number 8*, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1985.

Cunia, T. Construction of tree biomass tables by linear regression techniques. In: *Proceedings of the workshop on "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; 1986a.

Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: *Proceedings of the workshop on "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; 1986b.

Davies, O. L. *Statistical methods in research and production*. 3rd edition Oliver and Boyd, London; 1961.

AN OPTIMIZATION MODEL TO CALCULATE THE NUMBER OF SAMPLE TREES AND PLOTS^{1/}

Tiberius Cunia

Professor of Statistics and Operations Research,
State University of New York College of Environ-
mental Science and Forestry, Syracuse, NY 13210

The error of the estimates of forest biomass inventory has two major components; one due to the sample of plots (where the trees are measured for diameter but not for biomass) and one due to the sample of trees (measured for both diameter and biomass) from which the biomass regression function was calculated. The size of the error is affected, among other things by (i) the size of the sample of plots and (ii) the size of the sample of trees. An approach is described for the calculation of optimum number of plots and trees to select so that either (i) minimum sampling costs are obtained for a given required precision in the estimate of the average biomass per acre or (ii) maximum precision is obtained when the costs of sampling are fixed.

Introduction

In a previous paper, Cunia (1986a) considers sampling designs for forest inventory consisting of two major phases; a first phase where the trees of sample plots are measured for diameter (but not for biomass) and a second phase where the trees are measured for diameter and biomass. The second phase trees provide a tree biomass regression function on diameter, which applied to the first phase trees yields estimates of the average biomass per acre. He has also presented a method to combine the error of the biomass regression of the second phase with that of the sample plots of the first phase, when the error of the average biomass per acre is calculated. His method assumes that (i) the sample plots and sample trees of the first and second phase respectively are selected independently of each other by simple random sampling and (ii) the regression function of biomass on diameter is linear and satisfies the usual assumptions of the weighted least squares method, in particular the assumption that the conditional variance of the tree biomass for given diameter is proportional to the fourth power of the tree diameter.

The procedure requires that the estimator w of the average biomass per acre μ be of the form

$$w = b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]'[z]$$

where $[b]$ is the estimator of the vector $[\beta]$ of coefficients of the regression function of tree biomass y on tree variables $x_1, x_2, \dots, x_m$, that is,

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

calculated from the data of the second phase sample, and $[z]$ is a vector of statistics calculated from the data of the first phase sample. Note that $[]$ and $[]'$ notation is used to denote vectors or matrices and transposed vectors or matrices respectively. It can be shown that, for the present case where w estimates the average biomass per acre, the statistics z are nothing but the estimators of the averages of the corresponding variables x expressed on a "per acre" basis.

For example, if the regression of tree biomass on diameter is of the parabolic form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = [\beta]'[x]$$

where $x_1 = 1$, x_2 = tree diameter and x_3 = squared tree diameter, then (i) $[b]$ is the estimator of $[\beta]$ calculated from the data of the second phase sample and (ii) $[z]$ is the estimator of the vector $[\mu_z]$ defined as

$$[\mu_z] = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \Sigma x_1 / \text{acre} \\ \Sigma x_2 / \text{acre} \\ \Sigma x_3 / \text{acre} \end{bmatrix} = \begin{bmatrix} \text{number of trees per acre} \\ \text{sum of tree diameters per acre} \\ \text{sum of squared diameters per acre} \end{bmatrix}$$

calculated from the first phase sample.

If the covariance matrices of $[b]$ and $[z]$ are denoted as $[\sigma_{bb}]$ and $[\sigma_{zz}]$ respectively, the variance of w is approximately equal to

$$\sigma_{ww} = [\beta]'[\sigma_{zz}][\beta] + [\mu_z][\sigma_{bb}][\mu_z]$$

If estimators $[S_{zz}]$ and $[S_{bb}]$ of these covariance matrices are calculated from the data of the first and second phase respectively, the variance of w can be estimated by the approximate formula

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

Expressed in this way, the variance of w can be viewed as having two additive variance components; the first $[b]'[S_{zz}][b]$ containing the error of the first phase sample, the second, $[z]'[S_{bb}][z]$ containing the error of the second sample.

The size of the first error component depends on the sample size of the first phase, among other things. Similarly, the size of the second error component depends on the size of the second phase sample. One of the most important problems in survey sampling is that of deciding how large the sample should be, in our case, how

^{1/}Paper based on a set of lecture notes "On the error of biomass studies in forest inventories; Part 1: Its major components". Faculty of Forestry Miscellaneous Publication Number 8 (ESF 85-004). SUNY College of Environmental Science and Forestry, Syracuse, NY.

large the samples of the first and second phase should be. Samples that are too large in size may yield estimates that are too precise for the needs of the management and, thus, the sampling process may be too wasteful of one's resources. On the other hand, small samples may lead to poor management decisions as the inventory estimates may lack sufficient precision; and this may also prove to be too costly. What one should do is to determine the size of the first and second phase samples that is expected to minimize the sum of (1) sampling costs and (2) losses due to management decisions based on inventory estimates.

Viewed this way, the problem may become too complex to solve, since a management loss function may be extremely difficult, if at all possible to derive. What one can do instead, is to approach the problem in the following way. Find the sample size that would either (i) minimize the sampling costs for required precision in the estimates of interest (in our case the estimate of the average biomass per acre), or (ii) maximize the precision (or minimize the error) of these estimates for given allowable costs of sampling. These two problems are equivalent, since the solution of one can be generally derived from the solution of the other.

A General Approach to the Optimization Process

To define a somewhat general optimization approach we shall start with the following assumptions.

(1) The variance V of the estimator of interest can be split into two additive components V_1 and V_2 associated with the samples of the first and second phase respectively.

(2) The variance component V_1 can be written as an explicit function of the various sizes $n_{11}, n_{12}, \dots$, of the first phase and, similarly, the variance component V_2 can also be written as a function of the various sizes $n_{21}, n_{22}, \dots$ of the second phase sample. Let us write these components as

$$V_1(n_{11}, n_{12}, \dots) \text{ and } V_2(n_{21}, n_{22}, \dots)$$

(3) The sampling costs of the first and second phase can be expressed, at least approximately as the cost function

$$C_1(n_{11}, n_{12}, \dots) \text{ and } C_2(n_{21}, n_{22}, \dots)$$

respectively, and the total sampling costs as their sum

$$C = C_1 + C_2$$

Then the problem of calculation of optimum sample size can be expressed as the following problem.

Mathematical programming problem: Find the values $n_{11}, n_{12}, \dots, n_{21}, n_{22}, \dots$ that minimize the cost function

$$C = C_1(n_{11}, n_{12}, \dots) + C_2(n_{21}, n_{22}, \dots)$$

subject to

$$V_1(n_{11}, n_{12}, \dots) + V_2(n_{21}, n_{22}, \dots) = V^*$$

where V^* is the required variance of the estimator of interest.

This is the problem of finding the sample size that yields required precision at minimum sampling costs. The equivalent problem of finding the sample size that optimizes precision (minimizes error) for a given cost can be similarly expressed as follows.

Equivalent mathematical programming problem:

Find the values $n_{11}, n_{12}, \dots$, and $n_{21}, n_{22}, \dots$, that minimize the variance function

$$V = V_1(n_{11}, n_{12}, \dots) + V_2(n_{21}, n_{22}, \dots)$$

subject to

$$C_1(n_{11}, n_{12}, \dots) + C_2(n_{21}, n_{22}, \dots) = C^*$$

where C^* is the allowable cost of sampling.

There is no general solution to this general mathematical programming problem. However, if the functions V_1, V_2, C_1 and C_2 have certain properties, one can use calculus methods of optimization and find the optimum solution. Such is the case when the functions V and C are continuous and have at least the first and second derivatives. Then, the optimum solution is a solution of the following system of simultaneous equations

$$\begin{aligned} \partial L / \partial n_{ij} &= \partial C_1 / \partial n_{ij} + \partial C_2 / \partial n_{ij} + \lambda \partial V_1 / \partial n_{ij} \\ &\quad + \lambda \partial V_2 / \partial n_{ij} = 0 \\ &\text{for all } i = 1, 2, \dots \text{ and } j = 1, 2, \dots \end{aligned}$$

$$\text{and } V_1 + V_2 = V^*,$$

where L is the Lagrangian function

$$L = C_1 + C_2 + \lambda(V_1 + V_2 - V^*)$$

It can be shown that one, conveniently selected solution of this system of simultaneous equations (with unknowns $n_{11}, n_{12}, \dots, n_{21}, n_{22}, \dots, \lambda$) is the optimum solution we seek, provided some additional conditions involving the second order derivatives (not mentioned here) are satisfied. These additional conditions are usually satisfied, but to solve the system of equations may require numerical methods. We shall now show that, in our case the solution can be expressed in a closed form as a set of formulae.

A Sample Size Optimization Model

Let us consider the oversimplified example of a forest area containing trees of a single species where (i) in the first phase n_p plots (or points) are selected at random and all their trees are measured for their diameter d at breast height, (ii) in the second phase n_t trees are selected at random and measured for their biomass

y and diameter d, (iii) the regression of tree biomass on diameter is assumed to be of the linear form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

where $x_1 = 1$ and the other variables x are functions of diameter and (iv) the average biomass per acre is estimated by the statistic

$$w = b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]'[z]$$

where $[b]$ is the weighted least squares estimator of $[\beta]$ calculated from the phase 2 data and $[z]$ is the estimator of the vector $[\mu_z]$ of the arithmetic means $\mu_1, \mu_2, \dots, \mu_m$ of $x_1, x_2, \dots, x_m$ respectively expressed on a "per acre" basis, calculated from the data of phase 1.

It has been stated above that the variance $V = \sigma_{ww}$ of w can be split into two additive components

$$V_1 = [\beta]'[\sigma_{zz}][\beta]$$

which can be associated with the error of the first phase sample and

$$V_2 = [\mu_z]'[\sigma_{bb}][\mu_z]$$

which can be associated with the error of the biomass regression of the second phase. We shall now show how to (i) sort out in explicit terms the effect of the sample sizes n_p and n_t in the two variance components V_1 and V_2 , (ii) construct a sampling cost function, (iii) construct an optimization model for the calculation of the sample size and (iv) solve the model and express the optimum solution as a set of formulae.

Error Component Due to the Sample Plots of Phase 1

Under the assumption that the sample plots (or points) of phase 1 are selected completely at random (and with replacement), the covariance of the z_i and z_j variables can be written as σ_{ij}/n_p , where σ_{ij} is the covariance of the plot (or point) variables s_i and s_j . The variable s_i is defined as the variable x_i expressed on a per acre basis, that is, $s_i = (\sum x_i/a)$ where "a" is the plot area and $\sum$ is taken over all the trees in the given plot. For the definition of s_i when we have a Bitterlich sample point, the reader is referred to Cunia (1985). For example, if the plot area is one-fifth of an acre and $x_1 = d$ (tree diameter) then $s_1 = (\text{sum of the tree diameters in the plot})/(.20)$. Consequently, the covariance matrix of $[z]$ can be written as

$$\sigma_{zz} = [\sigma]/n_p$$

where the ij -th element of $[\sigma]$ is the covariance of z_i and z_j and the variance component V_1 can be written as

$$V_1 = [\beta]'[\sigma][\beta]/n_p = Q_p/n_p$$

Error Component Due to the Biomass Regression of Phase 2

Under the assumption that the trees of the second phase sample are selected completely at random and under the other usual assumptions of the weighted least squares method of regression, the covariance matrix of $[b]$ can be written as approximately equal, on the average, to

$$[\sigma_{bb}] = [\theta]/n_t$$

where the definition of the matrix $[\theta]$ is relatively more complex. To better see the meaning of $[\theta]$ let us have a closer look at the method of weighted least squares as described by Cunia (1986b). The sample covariance matrix of $[b]$, expressed as $[S_{bb}] = S_{uu|v}[T]^{-1}$, is an estimate of the true but unknown covariance matrix

$$[\sigma_{bb}] = \sigma_{uu|v}[T]^{-1} = \sigma_{uu|v}[\bar{T}]^{-1}/n_t$$

where

$$[\bar{T}] = [T]/n_t$$

Because the ij -th element of $[T]$ is the sum of the cross-products of the (transformed) variables v_i and v_j , the corresponding ij -th element of $[\bar{T}]$ is the average cross-product $v_i v_j$. More formally, we can write

$$[\bar{T}] = \begin{bmatrix} \sum v_1^2/n_t & \sum v_1 v_2/n_t & \dots & \sum v_1 v_m/n_t \\ \sum v_1 v_2/n_t & \sum v_2^2/n_t & \dots & \sum v_2 v_m/n_t \\ \vdots & \vdots & \ddots & \vdots \\ \sum v_1 v_m/n_t & \sum v_2 v_m/n_t & \dots & \sum v_m^2/n_t \end{bmatrix}$$

Because the trees are selected completely at random, the ij -th element of $[\bar{T}]$ estimates the population mean value (expected value) of the cross-product $v_i v_j$. We shall define now the matrix $[T_E]$ as the matrix of the expected values of the cross-products $v_i v_j$ and the matrix $[\theta]$ as the product $\sigma_{uu|v}[T_E]^{-1}$. Of course, for a sample of a given size n_t , we do not obtain $[\theta]$ but a statistic $S_{uu|v}[\bar{T}]^{-1}$. This is an estimate of the matrix $\sigma_{uu|v}[T]^{-1}$ which in turn is an estimate of $\sigma_{uu|v}[T_E]^{-1} = [\theta]$.

Consequently, the variance component V_2 associated with the error of the biomass regression of the second phase can be written as

$$V_2 = [\mu_z]'[\theta][\mu_z]/n_t = Q_t/n_t$$

Cost Function

To optimize the sample size we need also take into account the sampling costs. Although the average costs of selecting one sample plot of the first phase and measuring its trees may be thought of as a function of sample size (the more plots one has to select and measure, the smaller the average sampling costs per plot tend to be, since the average distance between plots and the

travelling costs per plot tend to diminish) it may be reasonable to ignore the fixed costs of sampling and define, with an acceptable level of approximation, the cost function for the first phase sample as

$$C_1 = c_p n_p$$

where c_p = average cost of selecting, travelling and measuring a sample plot. Similar arguments lead to the selection of the cost function of the second phase as

$$C_2 = c_t n_t$$

where c_t = average cost of selecting, travelling and measuring a sample tree.

Consequently, the costs of sampling that are affected by the sample sizes of the first and second phase can be written as

$$C = C_1 + C_2 = c_p n_p + c_t n_t$$

Optimization Model

The mathematical programming problem can now be expressed as the following model

Optimization model. Find the sample sizes n_p and n_t that minimize the cost function

$$C = C_1 + C_2 = c_p n_p + c_t n_t$$

subject to

$$V_1 + V_2 = Q_p/n_p + Q_t/n_t = V^*$$

where

$$Q_p = [\beta]'[\sigma][\beta],$$

$$Q_t = [\mu_z]'[\theta][\mu_z], \text{ and}$$

V^* = required precision (expressed as the variance) of the estimator w of the average biomass per acre μ .

To find the optimum solution we write first the Lagrangian function

$$L = c_p n_p + c_t n_t + \lambda (Q_p/n_p + Q_t/n_t - V^*)$$

By taking the partial derivatives of L with respect to n_p , n_t and λ we obtain a system of three equations in three unknowns which solved, yield the optimum solution

$$n_p = (\sqrt{Q_p/n_p}) (\sqrt{c_p Q_p} + \sqrt{c_t Q_t}) / V^*$$

$$n_t = (\sqrt{Q_t/n_t}) (\sqrt{c_p Q_p} + \sqrt{c_t Q_t}) / V^*$$

and the value of λ is of no interest. These are the sample sizes that minimize the sampling costs for the required error V^* in the estimate w of the mean biomass per acre μ .

One may wish to find the sample sizes that maximize the precision (minimize the error V^*) when the sampling costs are given. Then, he may work with the mathematical programming problem

expressed as the following equivalent model.

Equivalent optimization model. Find the sample sizes n_p and n_t that minimize the variance function

$$V = Q_p/n_p + Q_t/n_t$$

subject to

$$c_p n_p + c_t n_t = C^*$$

where

C^* = allowable costs of sampling.

The optimum solution of this model is

$$n_p = C^* (\sqrt{Q_p/n_p}) / (\sqrt{c_p Q_p} + \sqrt{c_t Q_t})$$

$$n_t = C^* (\sqrt{Q_t/n_t}) / (\sqrt{c_p Q_p} + \sqrt{c_t Q_t})$$

Of course, the optimum values n_p and n_t we select must be positive integers and, thus, the nearest integer close to the values given by the above formulae are usually selected. It is implicitly understood that V^* and C^* are such that the resulting sample size n_t is sufficiently large to allow the calculation of the biomass regression function. For more details about the derivation of the above optimum solutions, the interested reader is referred to Cunia (1985).

That these two sets of solutions are equivalent can be seen from the fact that the ratio n_p/n_t is the same for both problems, that is

$$n_p/n_t = \sqrt{c_t Q_p / c_p Q_t}$$

This allows one to go from one to the other solution by a simple and straightforward procedure. For example, assume that to obtain an estimate with a required variance V^* one finds the optimum sample sizes n_p and n_t for the sample plots and sample trees respectively. This yields an estimated cost of sampling of $C^* = c_p n_p + c_t n_t$. Suppose now that C^* appears to be prohibitively high and the management decided that it cannot spend more than $C^{**} < C^*$. Then, the new optimum sample sizes n^{**} and n^{**} that minimize the error of the estimate for the given allowable cost of sampling of C^{**} can be determined by the formulae

$$n_p^{**} = n_p^* C^{**} / C^*$$

and

$$n_t^{**} = n_t^* C^{**} / C^*$$

To calculate these optimum sample sizes, one must have prior estimates of $[\beta]$, $[\mu_z]$, $[\sigma]$, $[\theta]$, c_p and c_t . This is seldom if ever the case. Furthermore, we do not know how good these estimates should be before the sample sizes as calculated would be of any value. It is possible that small errors in the prior estimates of these parameters may critically affect the optimum sample size values. Good sensitivity studies are needed to show whether the approach suggested here has any practical value. There is no need for theoretical studies; simulation techniques

can be used with great advantage.

Consequently, the formulae above should be used with great care. Any results that one may obtain from their use should be carefully analyzed. If they seem quite different from what is expected from an intuitive point of view, chances are that the results are indeed of questionable value. There is one case, however, where the formulae can be used to advantage. For example, assume that data from sample plots and trees become available and one is in a position to analyze these data. He can get estimates of $[\beta]$, $[\mu_z]$, $[\sigma]$, $[\theta]$ and costs c_p and c_t . Then, these estimates can be used in the formulae above, and one can verify whether, in a very approximate way, the ratio

(actual number of plots/ actual number of trees)

is sufficiently close to the ratio of optimum sample sizes.

Let us now illustrate the application of the above procedure to a numerical example.

Example 1 - Calculate the optimum sample sizes n_p , the number of one-fifth acre sample plots of the first phase, and n_t , the number of trees of the second phase sample, when

(1) the sample plots of the first phase are to be selected by simple random sampling (with replacement), the sample trees of the second phase are to be selected by simple random sampling (with replacement) and the samples of the two phases are statistically independent,

(2) the average cost of selecting and measuring a sample plot is estimated as $c_p = \$120$, the average cost of selecting and measuring a sample tree is estimated as $c_t = \$50$ and the function of sampling costs is well approximated by the linear function

$$C = c_p n_p + c_t n_t$$

and (3) the regression function of the tree biomass y on the tree diameter d is of the parabolic form

$$\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = [\beta]'[x]$$

with the obvious definitions for x_1 , x_2 , x_3 , $[\beta]$ and $[x]$.

Assume that

(1) the estimates of the mean vector and the covariance matrix of the plot variables

s_1 = (number of trees in the plot) per acre

s_2 = (sum of tree diameters in the plot) per acre, and

s_3 = (sum of squared tree diameters in the plot) per acre

where s_i (defined as x_i expressed on a per

acre basis, $i = 1, 2, 3$) are those calculated in Example 1, of the earlier paper by Cunia (1986a) that is

$$[\mu_z] = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 123.69330 \\ 1090.6982 \\ 11380.879 \end{bmatrix}$$

and

$$[\sigma] = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{12} & s_{22} & s_{23} \\ s_{13} & s_{22} & s_{33} \end{bmatrix} = \begin{bmatrix} 10390.237 & 89963.308 & 863918.12 \\ 89963.308 & 830078.46 & 8622243.4 \\ 863918.12 & 8622243.4 & 98423671 \end{bmatrix}$$

and

(2) the estimates of the vector $[\beta]$ of regression coefficients and the matrix $[\theta] = \sigma_{uu|v}[\bar{T}]^{-1}$, are those given in the same Example 1, that is,

$$[\beta] = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 5.1818118 \\ -25.653078 \\ 12.988357 \end{bmatrix}$$

and

$$[\theta] = \sigma_{uu|v} [\bar{T}]^{-1} = n_t [s_{bb}] = \begin{bmatrix} 3076707.6 & -784538.32 & 45431.073 \\ -784538.32 & 205444.48 & -12276.279 \\ 45431.073 & -12276.279 & 767.58376 \end{bmatrix}$$

since $n_t = 353$ and

$$[s_{bb}] = \begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

Then, calculate the optimum future sample sizes n_p and n_t such that

(1) the minimum costs C of selecting and measuring n_p and n_t trees are obtained when the desired variance of w , the estimator of the average biomass per acre, is $\sigma_{ww} = 20,000,000$ (pounds)², and

(2) the optimum precision (minimum variance) of w , the estimator of the average biomass per acre μ is obtained for a maximum allowable cost of sampling of $C = \$120,000$ of selecting and measuring n_p plots and n_t trees.

Also, show that

(3) the solutions of questions (1) and (2) above are equivalent, that is, one can go from one to the other problem solution by simple calculations.

To solve these problems, we shall use the formulae of the present section and some of the

calculations already performed in Example 1 of Cunia (1986a)

(1) The optimum sample sizes n_p and n_t that minimize costs for desired precision are given by the formulae

$$n_p = (\sqrt{Q_p/c_p}) (\sqrt{c_p Q_p} + \sqrt{c_t Q_t}) / V^*$$

$$n_t = (\sqrt{Q_t/c_t}) (\sqrt{c_p Q_p} + \sqrt{c_t Q_t}) / V^*$$

where

$$c_p = 120, c_t = 50, V^* = \sigma_{ww} = 20,000,000$$

$$Q_p = [\beta]' [\sigma] [\beta] = 11497014000$$

$$Q_t = [\mu_z]' [\theta] [\mu_z] = 2345255900$$

Note that Q_p and Q_t can be calculated directly from the estimates of $[\beta]$, $[\mu_z]$, $[\sigma]$ and $[\theta]$ assumed above. But because of the identities

$$Q_p = (926) ([b]' [S_{zz}] [b] \text{ of Example 1 of Cunia (1986a)})$$

$$= (926) (12415782) = 11497014000$$

and

$$Q_t = (353) ([z]' [S_{bb}] [z] \text{ of Example 1 of Cunia (1986a)})$$

$$= (353) (6643784.4) = 2345255900$$

we have used the calculations already performed there.

As the intermediate results are

$$\sqrt{c_p Q_p} = \sqrt{(120) (11497014000)} = 1174581.5$$

$$\sqrt{c_t Q_t} = \sqrt{(50) (2345255900)} = 342436.6$$

$$(\sqrt{c_p Q_p} + \sqrt{c_t Q_t}) = 1174581.5 + 342436.6 = 1517018.1$$

$$\sqrt{Q_p/c_p} = \sqrt{11497014000/120} = 9788.1790$$

and

$$\sqrt{Q_t/c_t} = \sqrt{2345255900/50} = 6848.7311$$

we calculate the optimum sample sizes n_p and n_t as

$$n_p = (9788.1790) (1517018.1) / (20,000,000) = 742.44 = 742$$

and

$$n_t = (6848.7311) (1517018.1) / (20,000,000) = 519.48 = 519$$

(2) The optimum sample sizes n_p and n_t that maximize precision for the allowable sampling costs of \$120,000 are given by the formulae

$$n_p = C^* (\sqrt{Q_p/c_p}) / (\sqrt{c_p Q_p} + \sqrt{c_t Q_t})$$

and

$$n_t = C^* (\sqrt{Q_t/c_t}) / (\sqrt{c_p Q_p} + \sqrt{c_t Q_t}),$$

where

$C^* = 120,000$ and all other intermediate results have already been calculated in (1) above. Then,

$$n_p = (120,000) (9788.1790) / (1517018.1) = 774.27 = 774$$

and

$$n_t = (120000) (6848.7311) / (1517018.1) = 541.75 = 542$$

(3) The reader can verify that the solutions in (1) and (2) above are equivalent by calculating the ratios of the sample sizes n_p/n_t . In both cases, the ratio is approximately 1.4292. If one wants to go from the solution of (1) to the solution of (2), he proceeds as follows

Step 1 - The sampling costs in (1) above are

$$(742.44) (120) + (519.48) (50) = 115067$$

Step 2 - The sample sizes of (2) above are calculated as

$$n_p = (120000/115067) (742.44) = 774.27 = 774$$

and

$$n_t = (120000/115067) (519.48) = 541.75 = 542$$

Extension to the Case of More than One Species

The extension to more than one species is straightforward. Each species has a different biomass regression function and the interest lies in the estimation of the average biomass per unit area (acre) μ for all species combined. The procedure for calculating w , the estimator of μ and that of calculating an estimator of the variance of w is given by Cunia (1986a) for the case where (i) the first phase sample of plots and the second phase q samples of trees (one for each of the q species), are selected by simple random sampling method (ii) the $(q+1)$ samples of plots and trees are statistically independent and (iii) the biomass regression functions are linear. He shows that if

(1) $[b^1], [b^2], \dots, [b^q]$ represent the sample vectors of the regression coefficients of the q species and $[S_{bb}^{11}], [S_{bb}^{22}], \dots, [S_{bb}^{qq}]$ the estimators of their covariance matrices,

(2) $[z^1], [z^2], \dots, [z^q]$ represent the estimators of the first phase sample such that, $w_i = [b^i]' [z^i]$ defines the estimator of the mean biomass per unit area of species i , and $[S_{zz}^{ij}]$ is the estimator of the covariance matrix of $[z^i]$ and $[z^j]$, for $i, j = 1, 2, \dots, q$, and

(3) the $(q+1)$ samples of plots and trees are statistically independent of each other, that is, the covariance matrices $[S_{bb}^{ij}]$ and $[S_{zz}^{ij}]$ that is, of $[b^i]$ with $[b^j]$, and $[b^i]$ with $[z^j]$ respectively are all zero, then

$$(1) w = [B]' [Z] = w_1 + w_2 + \dots + w_q$$

= estimator of μ , the mean biomass per acre of all species 1, 2, ..., q combined, and

$$(2) S_{ww} = [B]' [S_{zz}] [B] + [Z]' [S_{bb}] [Z]$$

$$= [B]' [S_{zz}] [B] + [z^1]' [S_{bb}^{11}] [z^1]$$

$$+ [z^2]' [S_{bb}^{22}] [z^2] + \dots + [z^m]' [S_{bb}^{mm}] [z^m]$$

= estimator of the variance of w
where

$$[B]' = [[b^1]' \quad [b^2]' \quad \dots \quad [b^q]']$$

$$[Z]' = [[z^1]' \quad [z^2]' \quad \dots \quad [z^q]']$$

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [0] & \dots & [0] \\ [0] & [S_{bb}^{22}] & \dots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \dots & [S_{bb}^{qq}] \end{bmatrix}$$

and

$$[S_{ZZ}] = \begin{bmatrix} [S_{zz}^{11}] & [S_{zz}^{12}] & \dots & [S_{zz}^{1q}] \\ [S_{zz}^{21}] & [S_{zz}^{22}] & \dots & [S_{zz}^{2q}] \\ \vdots & \vdots & \ddots & \vdots \\ [S_{zz}^{q1}] & [S_{zz}^{q2}] & \dots & [S_{zz}^{qq}] \end{bmatrix}$$

Let us now give the formulae for the calculation of the optimum sample sizes

n_p = number of sample plots of the first phase, and

n_i = number of sample trees of species i of the second phase, for $i = 1, 2, \dots, q$,

when the $(q+1)$ samples of plots and trees are statistically independent, and they are all selected by simple random sampling (with replacement). We shall assume that

(1) The costs of sampling are sufficiently well approximated by the linear cost function

$$C = c_p n_p + c_1 n_1 + c_2 n_2 + \dots + c_q n_q$$

where n_p = average cost of selecting and measuring a sample plot

n_i = average cost of selecting and measuring a sample tree of species i , $i = 1, 2, \dots, q$

and C = (total sampling costs - fixed costs of sampling)

(2) The variance of the estimator w = $[B]'[Z]$ of μ , the average biomass per acre is given by a formula of the form

$$V = V_p + V_1 + V_2 + \dots + V_q$$

where

$$V_p = [B]'[\sigma][B]/n_p = Q_p/n_p$$

$$V_i = [\mu_z^i]'[\theta^{ii}][\mu_z^i]/n_i$$

$$= Q_i/n_i, \text{ for } i = 1, 2, \dots, q,$$

$$[B]' = [[b^1]' \quad [b^2]' \quad \dots \quad [b^q]']$$

$$[\mu_z^i] = \text{expected value of } [z^i]$$

and $[\sigma]$ and $[\theta^{ii}]$ are matrices of the type defined for one species in the previous section. Of course, matrix $[\sigma]$ contains the covariances of a much larger number of variables s and $[\theta^{ii}]$ refers to the matrix $[\theta] = \sigma_{uu} v[\bar{T}]^{-1}$ of the species i . Under these assumptions, Cunia (1965) gives the following solutions to two equivalent problems of optimization of future sample size.

Solution - The future sample sizes $n_p, n_1, n_2, \dots, n_q$ that minimize the cost function

$$C = c_p n_p + c_1 n_1 + c_2 n_2 + \dots + c_q n_q$$

subject to

$$V_p + V_1 + V_2 + \dots + V_q = Q_p/n_p + Q_1/n_1$$

$$+ Q_2/n_2 + \dots + Q_q/n_q = V^*$$

where V^* is the required precision (expressed as variance) of the estimator w can be calculated by the formulae

$$n_p = (\sqrt{Q_p/c_p}) (\sqrt{c_p Q_p} + \sqrt{c_1 Q_1} + \sqrt{c_2 Q_2} + \dots + \sqrt{c_q Q_q}) / V^*$$

and

$$n_i = (\sqrt{Q_i/c_i}) (\sqrt{c_p Q_p} + \sqrt{c_1 Q_1} + \sqrt{c_2 Q_2} + \dots + \sqrt{c_q Q_q}) / V^* \text{ for } i = 1, 2, \dots, q$$

Equivalent solution - The future sample sizes $n_p, n_1, n_2, \dots, n_q$ that minimize the variance of w

$$V = Q_p/n_p + Q_1/n_1 + Q_2/n_2 + \dots + Q_q/n_q$$

subject to

$$c_p n_p + c_1 n_1 + c_2 n_2 + \dots + c_q n_q = C^*$$

where C^* is a given value for allowable total cost of sampling, can be calculated by the formulae

$$n_p = C^* (\sqrt{Q_p/c_p}) / (\sqrt{c_p Q_p} + \sqrt{c_1 Q_1} + \sqrt{c_2 Q_2} + \dots + \sqrt{c_q Q_q})$$

and

$$n_i = C^* (\sqrt{Q_i/c_i}) / (\sqrt{c_p Q_p} + \sqrt{c_1 Q_1} + \sqrt{c_2 Q_2} + \dots + \sqrt{c_q Q_q}) \text{ for } i = 1, 2, \dots, q$$

As with the case of one species, to calculate the optimum sample sizes, one must have prior estimates of the parameters $[B]$, $[\mu_z^i]$, $[\sigma]$, $[\theta^{ii}]$, c_p , and c_i . This is seldom, if ever, the case. But if this is the case, then, one can use, from some previous sample data derived from some similar forest areas, the estimators $[B]$ of $[B]$, $[z^i]$ of $[\mu_z^i]$, $[S]$ of $[\sigma]$, and

$$Q_p = n_p [B]' [S_{ZZ}] [B]$$

$$Q_1 = n_1 [S_{bb}^{11}], Q_2 = n_2 [S_{bb}^{22}], \dots,$$

$$Q_q = n_q [S_{bb}^{qq}]$$

where $n_p, n_1, n_2, \dots, n_q$ are the sample sizes of the previous sample data and not the optimum sample sizes we seek to calculate.

Example 2 - Calculate the optimum sample sizes

n_p = number of one-fifth acre sample plots of the first phase

n_1 = number of trees of species group 1 of the second phase sample

n_2 = number of trees of species group 2 of the second phase sample

and

n_3 = number of trees of species group 3 of the second phase sample

when

(1) all four samples of plots and trees of various species are selected by simple random sampling (with replacement) with all four samples being statistically independent of each other,

(2) the average costs of selecting, measuring and processing a sample plot or tree of a given species are approximately equal to

$c_p = \$120, c_1 = \$30, c_2 = \$50, \text{ and } c_3 = \$60$

(3) the sampling costs function C is well approximated by the linear function

$$C = c_p n_p + c_1 n_1 + c_2 n_2 + c_3 n_3$$

and

(4) the regression functions of tree biomass y on tree diameter d are of the parabolic form

$$\hat{y} = \beta_{i1} + \beta_{i2}d + \beta_{i3}d^2 = \beta_{i1}x_1 + \beta_{i2}x_2 + \beta_{i3}x_3 = [\beta^i]'[x]$$

for species group $i = 1, 2, 3$ where the definition of x_j and vectors $[\beta^i]$ and $[x]$ is obvious.

Assume the following

(1) Rough estimates of the mean vector $[\mu_s]$ (which is the same as $[\mu_z]$) and covariance matrix $[\sigma]$ of the plot variables $s_1, s_2, \dots, s_9$ defined as

s_1 = (number of trees of species group 1 in a given plot) per acre

s_2 = (sum of diameters of the trees of species 1 in a given plot) per acre

s_3 = (sum of squared diameters of the trees of species 1 in a given plot) per acre

with similar definitions for s_4, s_5, s_6 (of species group 2) and s_7, s_8, s_9 (for species group 3) are those calculated in Example 2 of Cunia (1986a), that is,

$$[\sigma] = \begin{bmatrix} [\sigma^{11}] & [\sigma^{12}] & [\sigma^{13}] \\ [\sigma^{21}] & [\sigma^{22}] & [\sigma^{23}] \\ [\sigma^{31}] & [\sigma^{32}] & [\sigma^{33}] \end{bmatrix}$$

and

$$[\mu_z] = \begin{bmatrix} [\mu_z^1] \\ [\mu_z^2] \\ [\mu_z^3] \end{bmatrix} = \begin{bmatrix} 28.439525 \\ 245.73704 \\ 2592.1076 \\ 39.994600 \\ 352.64039 \\ 3655.4591 \\ 55.259179 \\ 492.32073 \\ 5133.3124 \end{bmatrix}$$

where

$$[\sigma^{ij}] = (926)[S^{ij}]_{zz} \text{ of Example 2 of Cunia (1986a), } i, j = 1, 2, 3$$

For example,

$$[\sigma^{11}] = (926) \begin{bmatrix} 3.3631663 & 30.013376 & 299.99303 \\ 30.013376 & 278.99377 & 2924.2532 \\ 299.99303 & 2924.2532 & 32478.880 \end{bmatrix}$$

$$= \begin{bmatrix} 3114.2920 & 27792.387 & 277793.55 \\ 27792.387 & 258348.23 & 2707858.4 \\ 277793.55 & 2707858.4 & 30075443 \end{bmatrix}$$

The multiplication of $[S^{ij}]$ by 926 = n_p was necessary because $[S^{ij}]$ represented the covariance matrix of $[z^i]$ with $[z^j]$, the averages of the 926 plot values $[s^i]$ and $[s^j]$.

(2) Rough estimates of the vectors $[\beta^i]$ of regression coefficients and matrices $[\theta^{ii}]$ those calculated in Cunia (1986a), that is

$$[\beta^1] = \begin{bmatrix} 295.60183 \\ -107.06967 \\ 16.882552 \end{bmatrix}, [\beta^2] = \begin{bmatrix} -256.70604 \\ 40.050701 \\ 9.1695394 \end{bmatrix}$$

$$[\beta^3] = \begin{bmatrix} 18.800242 \\ -20.693393 \\ 13.156786 \end{bmatrix}$$

and $[\theta^{ii}] = S_{uuly}[\bar{T}]^{-1} = n_i[S^{ii}]_{bb}$ of Cunia (1986a), $i = 1, 2, 3$, with

$$[\theta^{11}] = (100) \begin{bmatrix} 13256.622 & -3494.4363 & 211.95740 \\ -3494.4363 & 943.83938 & -58.826156 \\ 211.95740 & -58.826156 & 3.8046237 \end{bmatrix}$$

$$= \begin{bmatrix} 1325662.2 & -349443.63 & 21195.740 \\ -349443.63 & 94383.938 & -5882.6156 \\ 21195.740 & -5882.6156 & 380.46237 \end{bmatrix}$$

and similarly

$$[\theta^{22}] = (107) \begin{bmatrix} 25911.067 & -6691.9889 & 393.93612 \\ -6691.9889 & 1777.8435 & -108.29003 \\ 393.93612 & -108.29003 & 6.9277171 \end{bmatrix}$$

and

$$[\theta^{33}] = (146) \begin{bmatrix} 25197.692 & -6243.4168 & 347.66044 \\ -6243.4168 & 1587.0539 & -91.114812 \\ 347.66044 & -91.114812 & 5.4758847 \end{bmatrix}$$

We shall calculate the optimum sample sizes for the two cases of (1) minimizing the sampling costs for a required variance of w of approximately equal to $V^* = 18129960$ (pounds)² (same precision as that obtained in Cunia (1986a)) and (2) minimizing the variance of estimator w for the allowable sampling costs of $C^* = \$128230$ (the same costs of sampling as those of Cunia (1986a)). For this we need estimates of Q_p , Q_1 , Q_2 , ..., Q_q as well as estimates of functions of the form $\sqrt{Q_p/c_p}$, $\sqrt{Q_1/c_1}$, $\sqrt{c_p Q_p}$, etc. These estimates are calculated as follows, where many of the intermediate results are read from Cunia (1986a).

$$Q_p = [B]'[\sigma][B] = (926)[B]'[S_{zz}][B] \\ = (926)(12145981) = 11247179000$$

$$Q_1 = [z^1]'[\theta^{11}][z^1] = (100)[z^1]'[S_{bb}^{11}][z^1] \\ = (100)(746466.01) = 74646601$$

$$Q_2 = (107)(2338326.6) = 250200950$$

$$Q_3 = (147)(2899185.7) = 426180290$$

$$\sqrt{Q_p/c_p} = \sqrt{11247179000/120} = 9681.2442$$

$$\sqrt{Q_1/c_1} = \sqrt{74646601/30} = 1577.4093$$

$$\sqrt{Q_2/c_2} = \sqrt{250200950/50} = 2236.9665$$

$$\sqrt{Q_3/c_3} = \sqrt{426180290/60} = 2665.1463$$

$$\sqrt{c_p Q_p} = \sqrt{(120)(11247179000)} = 1161749.3$$

$$\sqrt{c_1 Q_1} = \sqrt{(30)(74646601)} = 47322.28$$

$$\sqrt{c_2 Q_2} = \sqrt{(50)(250200950)} = 111848.32$$

$$\sqrt{c_3 Q_3} = \sqrt{(60)(426180290)} = 159908.78$$

and

$$\sqrt{c_p Q_p} + \sqrt{c_1 Q_1} + \sqrt{c_2 Q_2} + \sqrt{c_3 Q_3} = 1480828.7$$

Consequently, we are now ready to apply the formulae of this section and find

(1) The optimum sample sizes that minimize the sampling costs for a required precision of $V^* = 18129960$

$$n_p = (9681.2442)(1480828.7)/(18129960) = 791$$

$$n_1 = (1577.4093)(1480828.7)/(18129960) = 129$$

$$n_2 = (2236.9665)(1480828.7)/(18129960) = 183$$

$$n_3 = (2665.1463)(1480828.7)/(18129960) = 218$$

(2) The optimum sample sizes that minimize the variance of w for allowable costs of sampling of $C^* = \$128230$

$$n_p = (128230)(9681.2442)/(1480828.7) = 838$$

$$n_1 = (128230)(1577.4093)/(1480828.7) = 137$$

$$n_2 = (128230)(2236.9965)/(1480828.7) = 194$$

$$n_3 = (128230)(2665.1463)/(1480828.7) = 231$$

Note that the precision and costs of sampling of the data of Example 2 of Cunia (1986a) are equal to $V = 18129660$ (pounds)² and $C = \$128230$ respectively. The first set of optimum sample sizes $n_p = 791$, $n_1 = 129$, $n_2 = 183$, and $n_3 = 218$ are expected to yield the same precision V at the minimum sampling costs of

$$C = (791)(120) + (129)(30) + (183)(50) + (218)(60) = 121020$$

for a saving of $\$(128230 - 121020) = \$7210$ or about $(100)(7210)/(128230) = 5.6$ percent. Similarly the second set of optimum sample sizes $n_p = 838$, $n_1 = 137$, $n_2 = 194$, $n_3 = 231$ minimizes the variance of the estimator w for the same sampling costs of the data of Example 2 of Cunia 1986a. The expected precision is about

$$V = Q_p/n_p + Q_1/n_1 + Q_2/n_2 + Q_3/n_3 \\ = 11247179000/838 + 74646601/137 \\ + 250200950/194 + 426180290/231 \\ = 13421455 + 544865 + 1289696 + 1844936 \\ = 17100952$$

or, a reduction in variance of about

$$(100)(18129660 - 17100952)/(18129660) \\ = 5.6 \text{ percent}$$

the same percent we have found with the first set of sample sizes. This was expected.

Of course, all these statements are strictly correct only if the assumptions made about the values of the variance and costs parameters are also strictly correct. As this is not the case, some sensitivity analysis should be made if one desires to see by how much precision and costs are affected by our assumptions. This is not being done here. From a practical point of view, however, we can say that the actual sample sizes of Example 2 are not too far from optimum, since the optimum sizes are expected to improve the efficiency only by less than 6 percent.

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

- Cunia, T. On the error of biomass estimates in forest inventories: Part 1: Its major components. Faculty of Forestry Miscellaneous Publication Number 8, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1985.
- Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the

workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Env. Sc. & Forestry, Syracuse, NY; 1986a.

Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Env. Sc. & Forestry, Syracuse, NY; 1986b.

TUTORIAL PAPERS

Error of Biomass Regressions

Tiberius Cunia

Professor of Statistics and Operations Research,
State University of New York College of Environ-
mental Science and Forestry, Syracuse, NY, 13210

The paper describes the weighted least squares method of linear regression and its application to the problem of estimating tree biomass regressions and their error. It discusses in detail the four main sources of error (tree selection, biomass measurement of sample trees, statistical model used and application of biomass regressions) and it shows how to calculate the three basic statistics of linear regression functions that are needed in forest inventory, namely [b], the estimate of the vector of regression coefficients, $S_{yy|x}$, the estimate of the conditional variance of y about the regression function and $[S_{bb}]$ the estimate of the covariance matrix of [b]. The procedure is then applied to a sample of trees to construct a biomass table on tree diameter, together with its 95 percent confidence and prediction intervals.

Introduction

In a previous paper, T. Cunia (1986a) proposed a procedure for combining the error of the tree biomass regression function with that from the sample plots, when the error of the forest inventory estimates is being calculated. This procedure requires that (i) the tree biomass regression function be of a linear form, that is

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

where $x_1 = 1, x_2, x_3, \dots, x_m$ are the independent variables and notation [] and []' is used to denote matrices (and vectors) and their transposes respectively and (ii) valid estimates [b] of [β] and $[S_{bb}]$ of the covariance matrix of [b] be available.

It is the objective of the present paper to discuss the problem of error of biomass regressions, identify its main sources and present the procedure for the estimation of this error by the method of least squares linear regression. It will be assumed here that the classical assump-

tions of this method are satisfied. For more details on the procedure of combining the error of the biomass regression with that of the sample plots the reader is referred to the above-mentioned paper.

Main Sources of Error

There are four major sources of error in tree biomass regression functions. The first is that associated with the selection of sample trees. The same selection procedure applied to the same tree population on different occasions results in different sets of sample trees and, thus, in different biomass regressions. The second source of error is the measurement of the sample trees. Different measurements of the same sample trees, made on different occasions by various individuals, do not ordinarily yield the same measurements. While the volume of the merchantable bole of a felled tree, or the green weight of the entire tree can be measured without any appreciable error, the measurement of the oven-dry biomass of a given tree component is usually based on subsampling; small parts of the tree component are selected by some random procedure, their dimensions are measured in the field and the value of their oven-dry biomass determined in the laboratory is used to estimate the biomass of the given tree component. The difference between the true, conceptual value of the tree attribute we want determined and the actual, recorded value obtained by measurement is known as measurement error. The third source of error is that of the statistical model used. Given the same set of sample tree data, different statisticians may use different models and, thus, obtain different biomass regressions. The basic assumptions of the model must be satisfied by both, sample and population, in order for the conclusions to be valid. Finally, the fourth source of error is associated with the application of the biomass regressions to a specific forest inventory. This error component may become extremely important when the regression is applied to a forest population that is very different from that for which it was estimated. Strictly speaking, the biomass regressions are never applied to the original population; the populations of trees are dynamic and change with time.

Let us now consider, in more detail, each source of error separately.

Sample Tree Selection

Because the trees of the forest cannot be all measured for biomass, one must rely on sampling. Any sample of trees, selected by any procedure, can be used to calculate a biomass regression. But only when the selection procedure is statistical and, properly applied, leading to a representative sample of the population of interest, can the resulting regressions be statistically valid.

We say that a sample is representative of a population of interest when the error of the bio-

^{1/}Based on the paper "On volume tables and their contribution to the error of forest inventory estimates" In: Forstliche Nationalinventuren in Europa (National Forest Inventory in Europe) - D. R. Pelz and T. Cunia, (Eds.) Mitteilungen der Abteilung Für Forstliche Biometrie Universität Freiburg, 7800 Freiburg I. Br. - Federal Republic of Germany.

mass regression function calculated from that sample and applicable to that population can be evaluated,, at least conceptually, in quantitative statistical terms. It is not necessary for the estimates of the error to be precise and unbiased; it suffices that these estimates be sufficiently good under the statistical model assumed and that the basic model assumptions be sufficiently well satisfied.

In general, a sample is representative if, for the given selection procedure, each element of the population has a known non-zero probability of selection. It is not necessary for the probability to be known in absolute terms. Knowing, for example, that the elements are selected with equal probability or probability proportional to a measure of tree size is sufficient. However, the probability must be non-zero. For example, if the biomass regression is to be applied to all forest trees in the area of interest, one should not limit the selection of sample trees from those that are dominant and codominant, healthy with no apparent defects, etc.

In addition to being representative, the selection procedure must also be such that the resulting samples could be subjected to valid statistical analysis. Furthermore, it must also be cost-efficient, that is, yield biomass regression functions of acceptable precision at reasonable costs. This implies that one should not use a selection procedure that yields representative samples but for which statistical techniques for a valid analysis have not been devised yet, nor should one use a procedure that is too expensive for all practical purposes.

A search of the forest biomass literature made by Cunia (1979a,b) showed that many authors did not state how the sample trees were selected and from what type of tree population. Or, when they did, the procedure was non-random in the statistical sense; the sample trees were selected from subjectively selected parts of the given forest area, trees of odd shapes or defoliated were discarded from the sample, etc. Finally, when the sampling method was indeed statistical, the biomass regressions were not properly estimated. Sample trees selected by cluster or stratified sampling were analyzed by the ordinary least or weighted least squares method as if they were selected by simple random sampling.

Biomass Measurement of Sample Trees

Once the sample trees are selected, they must be measured for biomass. The biomass is usually required by components; either by major components such as main bole, crown or stump-root system, or minor components such as wood or bark of dead branches, wood of live branches larger than 10 cm. of diameter, bark of merchantable bole, etc. All these components require prior and very precise definition. A search of the literature shows a wide variety of definitions. There is a great need for standardization of terminology, if one is to combine results from the analysis of sample data from various sources

or compare biomass regressions published by various authors.

Assuming that a tree component has been previously defined, it is relatively easy to measure the fresh weight; it can be weighed in one piece or in smaller sections. The only sources of error may be in the physical separation of the component from the tree (with the possible loss of small particles), the actual weighing by mechanical devices and the possible loss of some humidity (between time of harvesting and time of measurement). The measurement of the volume of the components like the main bole that have sections easily approximated by geometrical bodies is also relatively precise. Much more difficult is the problem of measuring the volume of components such as leaves, small roots or branches, etc. or the ovendry weight of any component of large size. In all these cases one must rely on subsampling.

Consequently, the value assigned to a sample tree is most of the time an estimate, not the true value of the tree biomass. As such, it has a subsampling bias and random error. Many of the techniques we know for estimating the tree biomass have an inherent statistical bias, which is ordinarily small if the subsampling is done properly and the sample size is sufficiently large. The random part of the error depends on the method of subsampling, technique of estimation and sample size. When the biomass regressions are calculated by the statistical least squares methods, the random part of the subsampling error is automatically taken into account. This in itself is not a problem. Assuming that the subsampling bias is small, the real problem is that of efficiency. Is it better to have a small sample of trees for which the biomass is precisely estimated, or it is better to have a large sample of trees for which the ovendry biomass estimates have low precision? One way to optimize the combined sizes of the sample of trees and the subsamples within the trees is given in Cunia (1986 b).

Statistical Models

Seldom, if ever the assumptions of a statistical model are satisfied. This in itself is not unusual; it is part of the whole process of solving real world problems by abstracting them first as mathematical models and by solving then these models. The solutions found for the models are assumed to apply equally well to the real world problems they represent.

The errors associated with the mathematical model used are generally small when the model fits well the real world problem. It may become quite large if the basic assumptions of the model are critically violated, either by the population of trees or by the sample drawn from this population. Because we are concerned here with a regression model, let us discuss the basic assumptions of the least squares method and the effect these assumptions have on biomass regressions when they are not satisfied. For a more detailed

discussion of this problem, the reader is referred to Cunia (1979a,b).

There is first the assumption that the true regression function is of the assumed form. Because one deals with finite populations of trees, this assumption is never fulfilled in the strict statistical sense. But this in itself is not a serious drawback. Provided one works only with forms that were shown to be good, and he has no need to extrapolate the application of the estimated regression beyond the sample data, the effect of this assumption is minimal. In particular it does not seem to matter much whether the form of the regression function is linear or non-linear. This is important in view of the controversy going on between biologists who prefer working with non-linear functions of the allometric form and statisticians who most of the time prefer working with linear regressions.

It is true that some biological arguments can be brought in favor of allometric functions; the increase in biomass of a growing tree is proportional to the biomass contained in the tree. But it is also true that the form of the regression function is somewhat affected and greatly obscured by the inherent variation of biomass values from tree to tree. Furthermore, mathematical arguments based on Taylor's Theorem prove that one can always find linear functions that can approximate as closely as desired many non-linear functions as long as this approximation refers to a finite range. Consequently, the decision on whether to use linear or non-linear regression functions must be based entirely on different considerations.

The linear functions have several main advantages. Most of the statistical theory has been developed for linear regressions only. The least squares method is well known, it is simple to apply and has been extended to cover cases such as piecewise linear functions, harmonization and additivity of regression functions of component parts of tree biomass, selection of sample trees by methods other than simple random sampling, etc. Another big advantage is that the error of the biomass regression can be expressed in a convenient form that makes the method to combine it with the error of forest inventory sample plots relatively easy and straightforward to apply. Finally, there is enough empirical evidence to suggest that linear or non-linear functions properly selected are equally good. The only disadvantage one can think of, is that the non-linear regression functions may be better to use when they are applied beyond the range of the sample data; they seem to yield better estimates of the biomass when extrapolated.

The second assumption that the conditional variance of the tree biomass is homogeneous is generally not critical; forest biometricians know how to modify the least squares techniques when this assumption is not satisfied. One way is to transform the variables. Another way is to use the weighted least rather than the least squares method. But the assumption may become critical when (i) one uses poor transformations or

(ii) no correction for bias is made when proper transformations are used or finally, when (iii) wrong weights are used with the weighted least squares method.

The third assumption that the conditional distribution of the tree biomass is normal (in the statistical sense) is not critical unless (i) the sample of trees is relatively small, say 10-15 trees, in which case the central limit theorem does not necessarily apply and the confidence intervals and the results of significance tests may not be valid, or (ii) prediction limits for the biomass of individual trees are desired, in which case the calculated limits will be symmetrical when it is well known that the conditional probability distribution of the tree biomass is highly skewed. Otherwise the regression results are all right; the shape of the conditional distribution is not needed (as an assumption) when the point estimates of the regression coefficients and their covariance matrix are calculated, and with sufficiently large samples, say, at least 20-25 trees, the inferences based on normal distribution (about tests of significance or confidence intervals) are valid, since the central limit theorem applies sufficiently well.

The last assumption is that of statistical independence among the biomass measurements of the sample trees. It is an important assumption because the sample trees are almost never selected completely at random to satisfy this assumption. It is much less expensive to select trees in clusters and much more efficient to use stratification by size (tree diameter) or geographical location (site quality class, forest type and age class, etc.). To see what may happen when the sample trees are selected by methods other than simple random sampling (that insures statistical independence), and the ordinary least or weighted least squares are still applied in their standard form to calculate biomass regressions, one may refer to the empirical results and the results obtained by simulated sampling reported by Cunia (1981, 1986c). For some of the modifications to make to the least squares method when the sample trees are selected by methods other than random sampling, the reader is referred to Cunia (1986d).

Application of Biomass Regressions

The error component associated with the application of the biomass regression functions to a specific forest area is practically zero, when the regressions were estimated from trees properly selected from that specific area. The possible error is negligibly small even though tree populations are dynamic in nature and change with time; and, thus, the population from which the sample trees were selected is no longer the same population when the resulting biomass regressions are applied. The error is also negligibly small when the regression was estimated from a population similar to that being inventoried. However, this error component may be quite important in size for the forest inventory

systems for which old biomass regressions are being used; regressions calculated a long time before, from subjectively selected trees. When this occurs it is impossible to estimate the size of this error component.

Least and Weighted Least Squares Linear Regression Method

Before describing the specific weighted least squares procedure as applied to biomass regression estimation, we shall briefly describe the ordinary least squares method as applied to the estimation of linear regression functions. This will introduce the main concepts and notation used here and the three basic formulae for the three main statistics that summarize all the information from the sample tree data that is usually needed in forest inventory.

Least Squares Method

The basic assumptions of the least squares linear regression method were discussed in the previous section. We shall now be more specific and state them more formally as follows

(1) The expected value of y (the dependent variable) for given values of $x_1, x_2, \dots, x_m$ (the independent variables) is of the linear form

$$E(y|x) = \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

= regression function of y on $[x]$

where y = biomass of some tree component

$x_1, x_2, \dots, x_m$ = tree characteristics other than biomass; such as, for example, tree diameter, height, species (dummy variable) etc. with $x_1 = 1$ being the variable that introduces a constant term (the intercept) in the regression function

$[x]' = [x_1 \ x_2 \ \dots \ x_m]$ = vector of the fixed variables x

$[\beta]' = [\beta_1 \ \beta_2 \ \dots \ \beta_m]$ = vector of the regression coefficients ,

and $[]'$ = notation used to denote transposed vectors or matrices $[]$.

(2) The conditional variance of y given $[x]$, denoted here as $\sigma_{yy|x}$ is homogeneous, that is, the variation of y about the regression function remains the same no matter what the given values $x_1, x_2, \dots, x_m$ are. More formally,

$$E(y - E(y|x))^2 = \sigma_{yy|x}, \text{ a constant value}$$

(3) The covariance of any two sample values y_i and y_j is equal to zero, that is, the random variables y_i and y_j are uncorrelated. More formally

$$E(y_i - E(y_i|x_i))(y_j - E(y_j|x_j)) = 0$$

(4) The conditional probability distribution of y given $[x]$ is normal. This assumption is needed only when null hypotheses are tested for significance or prediction and confidence

limits are calculated. It also implies that y_i and y_j are statistically independent since two normally distributed variables that are uncorrelated are also statistically independent.

There are also several implicit assumptions that are not mentioned here. For example, the fixed variables $x_1, x_2, \dots, x_m$ are measured without error, the number n of sample elements is larger than the number m of independent variables x , there are at least m distinct vectors $[x]$ in the sample, or the population from which the sample elements are selected is static, that is, does not change with time.

Let the values $y, x_1, x_2, \dots, x_m$ of the k -th sample element, $k = 1, 2, \dots, n$ be denoted as

$$y_k, x_{k1}, x_{k2}, \dots, x_{km}$$

To apply the least squares method we shall arrange these values in the two matrices $[X]$ and $[Y]$ of sample values,

$$[X] = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \quad \text{and} \quad [Y] = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The matrices of cross products are now defined as

$$[T] = [X]'[X] \text{ and } [P] = [X]'[Y]$$

Then, the three statistics that summarize the information from the set of n sample elements are

(1) $[b] = [T]^{-1}[P]$ = estimator of the vector of regression coefficients $[\beta]$

(2) $S_{yy|x} = ([Y]'[Y] - [b]'[P])/(n-m)$ = estimator of the conditional variance $\sigma_{yy|x}$ of y given $[x]$, and

(3) $[S_{bb}] = S_{yy|x} [T]^{-1}$ = estimator of the covariance matrix $[\sigma_{bb}]$ of $[b]$

where $[T]^{-1}$ is the inverse of the matrix $[T]$.

Weighted Least Squares Method as Applied to Biomass Regressions

The basic assumptions of the least squares linear regression are all the same except for the second assumption (about the variance of y) which is now changed to the new assumption that the conditional variance of y given $[x]$ is proportional to a^2 , a value that is known for all sample (and population) elements. Ordinarily, a^2 is a known function of the given independent variables $x_1, x_2, \dots, x_m$. More formally, this can be expressed as

$\sigma_{yy|x} = \sigma_{uu|v} a^2$
 $=$ conditional variance of y given $[x]$
 where
 $a =$ known value, and
 $\sigma_{uu|v} =$ an unknown constant value (factor of proportionality)
 $=$ conditional variance of the new transformed variable $u = (y/a)$ given $[v] = [x]/a$

To calculate the weighted least squares estimates $[b]$ of the vector of regression coefficients $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ we apply the least squares method to the new transformed variables

$$u = y/a, v_1 = x_1/a, v_2 = x_2/a, \dots, v_m = x_m/a,$$

that is, we apply the least squares method to the new regression function of u on $[v]$

$$\hat{u} = \beta_1 v_1 + \beta_2 v_2 + \dots + \beta_m v_m = [\beta]'[v]$$

More specifically, we start with the calculation of the matrices of the transformed values $[U]$ and $[V]$, whose elements are

$$u_k = y_k/a, v_{k1} = x_{k1}/a, v_{k2} = x_{k2}/a, \dots, v_{km} = x_{km}/a$$

We continue with the calculation of the matrices of cross products

$$[T] = [V]'[V] \text{ and } [P] = [V]'[U]$$

and the calculation of the three basic statistics

- (1) $[b] = [T]^{-1}[P] =$ weighted least squares estimator of the vector of regression coefficients $[\beta]$
- (2) $S_{uu|v} = ([U]'[U] - [b]'[P])/(n-m)$
 $=$ estimator of the conditional variance $\sigma_{uu|v}$ of u given $[v]$, and
- (3) $[S_{bb}] = S_{uu|v}[T]^{-1} =$ estimator of the covariance matrix $[\sigma_{bb}]$ of $[b]$

Note the following:

- (1) $S_{yy|x} = a^2 S_{uu|v} =$ estimator of the conditional variance of y given $[x]$, a value which varies from tree to tree according to the tree vector $[x]$
- (2) When y is the biomass of the tree bole and $[x]$ is defined as a function of the tree diameter d , empirical evidence shows that " a " is approximately equal to d^2 . As the tree basal area is $\pi d^2/4$, the new variable $u = y/d^2$ represents a value which is proportional to the variable "bole biomass per square inch of tree basal area".
- (3) When y is the biomass of the tree bole and $[x]$ is defined as a function of the tree diameter d and height h , empirical evidence shows that " a " is approximately equal to $d^2 h$. As the bole volume is proportional to $d^2 h$ (recall that

the function $y = b d^2 h$ can be used as a regression function of y on d and h), the new variable $u = y/d^2 h$ represents a value which is approximately proportional to the variable "biomass per unit of bole volume".

Testing Null Hypotheses about $[\beta]$

Sometimes it may be of interest to test the null hypothesis that some regression coefficients are equal to zero; that is, the corresponding variables x may be eliminated from the regression function without diminishing the goodness of the regression function to estimate y . This accomplishes two things, it simplifies the regression function and it reduces the error of the regression estimators.

Because the variables x are arranged in an arbitrary way, let us show how to test the null hypothesis that, for some $r < m$,

$$\beta_{r+1} = \beta_{r+2} = \dots = \beta_m = 0$$

The test requires the assumption that the conditional probability distribution of y given $[x]$ is normal and consists of the following steps

Step 1 - Calculate the regression sums of squares

$$CR_1SS = [b_1]'[P_1] \text{ and}$$

$$CR_2SS = [b_2]'[P_2]$$

of the Unrestricted Regression R_1

$$R_1: \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r + \beta_{r+1} x_{r+1} + \dots + \beta_m x_m$$

and the Restricted (under the null hypothesis) Regression R_2

$$R_2: \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r$$

Step 2 - Calculate the test statistic

$$F = (n-m)(CR_1SS - CR_2SS)/(m-r)U_1SS$$

where

$$U_1SS = [Y]'[Y] - CR_1SS$$

Step 3 - From a table of F -distribution with $(m-r)$ and $(n-m)$ degrees of freedom, find the critical value F_α , for some probability α of rejecting the null hypothesis when the null hypothesis is true. Then, apply the decision rule to accept the null hypothesis when $F < F_\alpha$ and reject it otherwise.

The above procedure has been written for the least squares method. When the weighted least squares method is being used, the same procedure applies, except that

$$\begin{aligned}
 [P_1] &= [V_1]'[U_1], \\
 [P_2] &= [V_2]'[U_2], \text{ and} \\
 U_1SS &= [U]'[U] - [b_1]'[P_1]
 \end{aligned}$$

There is an alternate test for the special case where $r = m-1$, that is, for testing the null

hypothesis that $\beta_i = 0$ for some $i = 1, 2, \dots, m$. Then, the test statistic is

$$t = b_i / \sqrt{S_{b_i b_i}}$$

where $S_{b_i b_i}$ is the estimator of the variance of

b_i provided by $[S_{bb}]$. The critical value t_α is taken from a table of t-distribution with $(n-m)$ degrees of freedom for some probability level α of rejecting the null hypothesis, when the null hypothesis is true. The null hypothesis is accepted when $-t_\alpha < t < t_\alpha$ or rejected otherwise.

Constructing $(1-\alpha)$ Confidence Intervals for β_i and $[\beta]$

The $(1-\alpha)$ confidence intervals for some regression coefficient β_i is calculated by the formula

$$b_i \pm t \sqrt{S_{b_i b_i}}$$

where

t = critical value of t read from a table of t-distribution with $(n-m)$ degrees of freedom and for $(1-\alpha)$ confidence level, and

$S_{b_i b_i}$ = estimator of the variance of b_i , the i -th diagonal element of $[S_{bb}]$.

Note that the inferences about the confidence interval of β_i are strictly valid only if the conditional probability distribution of y given $[x]$ is normal; they are acceptably good even when the distribution is not normal but the sample size is sufficiently large.

Sometimes it may be interesting to make inferences about a joint confidence interval for the entire vector $[\beta]$. It is not right to calculate the confidence intervals for each β_i separately and then combine them; they are not statistically independent. To calculate such a joint confidence interval we start by defining

$$\begin{aligned} SSBB &= [b - \beta]'[T][b - \beta] \\ &= \text{sum of squares associated with the} \\ &\quad \text{vector } [\beta] \text{ of regression coefficients.} \end{aligned}$$

Because $[\beta]$ is not known, this sum of squares cannot be calculated. However, it can be used to define the $(1-\alpha)$ joint confidence interval of $\beta_1, \beta_2, \dots, \beta_m$ as the set of values $[\beta]$ that satisfy the inequality

$$[b - \beta]'[T][b - \beta] < m F S_{yy|x}$$

for the least squares method, or

$$[b - \beta]'[T][b - \beta] < m F S_{uu|x}$$

for the weighted least squares method, where F is taken from a table of F-distribution with m and $(n-m)$ degrees of freedom at the desired $(1-\alpha)$ confidence level.

Note that the statements about the confidence intervals of the individual regression coefficients β_i or the joint confidence interval of $[\beta]$ are strictly valid only if the assumption about normality of the conditional probability distribution of y given $[x]$ for the least squares (or u for given $[v]$ for the weighted least squares) is satisfied. Of course, because of the central limit theorem, the confidence interval statements are approximately all right when the sample size is sufficiently large.

Constructing Confidence and Prediction Intervals Associated with the Regression Estimates

Assume that an element of the population is selected at random, its values $x_1, x_2, \dots, x_m$ are measured and, thus, they are known and we are interested in calculating point and $(1-\alpha)$ interval estimators for (i) the expected value of y for that given $[x]$ and (ii) the actual value of y for that particular element.

Let the measured values $x_1, x_2, \dots, x_m$ be written as the vector

$$[x_0]' = [x_{01} \ x_{02} \ \dots \ x_{0m}]$$

the expected value (the arithmetic mean) of all population elements that have $[x] = [x_0]$ be written as μ_0 and the actual value (a random variable) taken on by the selected tree be written as y_0 . Then

(1) the point estimator of both μ_0 and y_0 is the regression estimator

$$\hat{y}_0 = [b]'[x_0]$$

(2) the $(1-\alpha)$ confidence interval of the expected value of y given $[x] = [x_0]$ is given by

$$\hat{y}_0 \pm t \sqrt{S_{\hat{y}_0 \hat{y}_0}}$$

where

$$\begin{aligned} S_{\hat{y}_0 \hat{y}_0} &= [x_0]'[S_{bb}][x_0] \\ &= \text{estimator of the variance of } \hat{y}_0, \\ &\quad \text{that is, the variance of the estimation error } (\hat{y}_0 - \mu_0), \text{ and} \end{aligned}$$

t = value read from a table of t-distribution with $(n-m)$ degrees of freedom at the $(1-\alpha)$ confidence level.

(3) the $(1-\alpha)$ prediction interval of the actual value of y given $[x] = [x_0]$ is given by

$$\hat{y}_0 \pm t \sqrt{S_{y_0 y_0}}$$

where

$$\begin{aligned} S_{y_0 y_0} &= \hat{S}_{\hat{y}_0 \hat{y}_0} + S_{yy|x_0} \\ &= \text{estimator of the variance of} \\ &\quad \text{the prediction error } (y_0 - \hat{y}_0), \\ &\quad \text{and} \end{aligned}$$

t has been defined above

Note that, for the case of weighted least squares,

$$s_{yy|x_0} = a^2 s_{uu|v}$$

and, thus,

$$s_{y_0 y_0} = s_{uu|v} (a^2 + [x_0]' [T]^{-1} [x_0])$$

It may also be of interest here to state the basic difference between the confidence and prediction intervals; the confidence interval may or may not include the parameter μ_0 , a fixed value, while the prediction interval refers to an interval in which a random variable y_0 may or may not fall in, at some future time.

A Numerical Example

Let us use the data from 353 sample trees reported by Cunia (1985). These trees were selected by simple random sampling and their diameter at breast height (in inches) and total above ground biomass y (in pounds of green weight) are listed by species group in Table 1. To estimate the regression function of y on d , for all species combined, we must select a mathematical model to represent the regression function and the conditional variance of y given d .

From past experience it is known that, an acceptably good mathematical model to express the regression function of the tree biomass y on diameter d is that of the rectangular parabolae

$$\begin{aligned} y &= \beta_1 + \beta_2 d + \beta_3 d^2 \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = [B]' [x] \end{aligned}$$

where $x_1=1$, $x_2=d$ and $x_3=d^2$. This can also be seen from the plot of the values y over the values d of Table 1 shown in Figure 1. Also from past experience it is known that the conditional variance of y given d is approximately proportional to d^4 . That this is reasonable can be seen from Figure 2 where the conditional standard deviation of y given d is plotted against the squared diameter d^2 . Note that the statement "conditional variance of y given d is proportional to d^4 " is equivalent to the statement "conditional standard deviation of y given d is proportional to d^2 " and also equivalent to the statement "a straight line passing through the origin of the two axes" is a good expression of the relationship between conditional standard deviation of y given d and the squared diameter d^2 .

To calculate the three main statistics of the regression function by the weighted least squares method we proceed as follows. We start with the calculation of the new transformed variables

$$\begin{aligned} u_k &= y_k / d_k^2 \\ v_{k1} &= x_{k1} / d_k^2 = 1 / d_k^2 \\ v_{k2} &= x_{k2} / d_k^2 = d_k / d_k^2 = 1 / d_k \\ v_{k3} &= x_{k3} / d_k^2 = d_k^2 / d_k^2 = 1 \end{aligned}$$

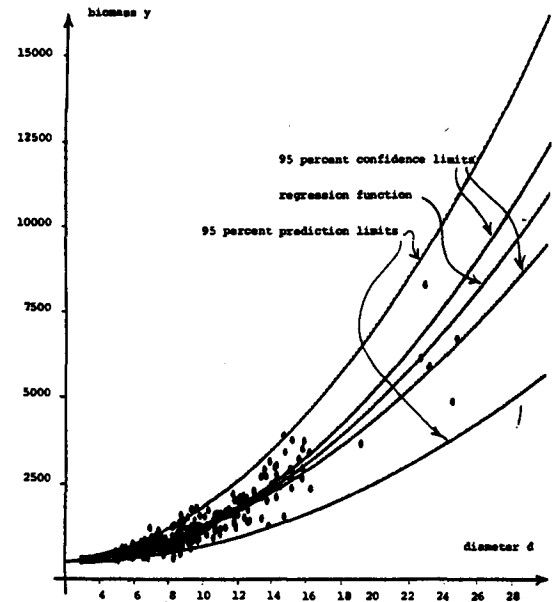


Figure 1 - Total above ground tree biomass y (green weight in pounds) plotted against tree diameter d (inches) together with the parabolic weighted least squares biomass regression and the corresponding 95 percent confidence and prediction limits.

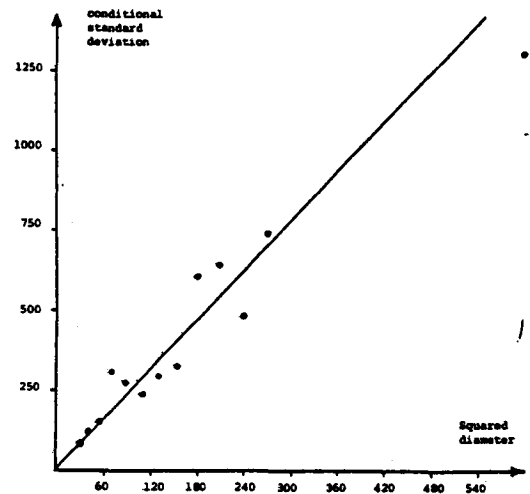


Figure 2 - The sample conditional standard deviation $s_{yy|x}$ of biomass y for given diameter d plotted against the squared diameter, and a straight line passing through the origin, an expression of the relationship between conditional standard deviation and squared diameter.

Table 1 - Diameters d (inches) and total above ground biomass y (pounds of green weight) of 353 trees arranged by species groups.

d	y	d	y	d	y	d	y	d	y	d	y	d	y	d	y
<u>Species group 1</u>															
9.3	557	7.8	621	9.5	813	8.2	546	7.2	322	9.3	647	12.7	1220	14.3	1953
11.4	1529	8.3	790	12.5	1356	8.7	627	12.0	1055	5.6	236	7.5	409	9.3	866
7.9	369	7.2	507	6.2	288	9.3	936	6.5	272	7.6	460	6.1	234	6.6	272
12.4	1993	5.0	206	7.2	356	7.5	436	7.8	533	5.0	123	9.4	419	7.1	503
6.6	336	5.1	240	11.1	851	6.6	400	11.4	1532	7.6	410	13.9	1125	9.2	944
8.4	637	5.7	235	9.8	821	9.5	846	6.5	306	8.5	497	5.5	221	6.3	276
11.9	1277	5.3	251	7.1	394	6.5	296	7.4	475	8.6	444	5.1	244	9.5	647
13.4	2539	7.1	513	5.3	178	9.0	771	14.2	2376	13.5	1315	5.4	275	9.2	1005
7.0	439	7.2	474	8.2	482	9.2	717	6.1	210	7.8	410	6.0	186	9.5	962
11.1	1471	5.1	219	8.6	626	11.7	1534	7.6	414	5.6	234	6.0	141	--	--
6.7	364	22.8	8195	7.7	594	15.8	2518	11.4	1057	6.3	268	5.5	187	--	--
7.0	377	8.0	339	9.5	756	11.9	1864	13.5	1956	9.5	885	5.1	158	--	--
6.9	384	8.1	577	6.8	378	6.9	388	5.2	149	5.8	225	5.2	170	--	--
<u>Species group 2</u>															
7.2	762	5.4	435	6.3	513	6.5	603	5.5	145	14.0	2311	7.0	469	8.8	841
6.8	485	8.7	1091	6.6	355	16.3	2219	8.0	383	8.9	457	6.6	327	7.7	529
6.3	322	6.4	505	5.0	250	5.2	175	12.0	1124	13.8	1831	15.2	2531	11.6	1607
6.1	210	8.7	1186	10.4	1563	11.9	1190	6.2	278	16.2	3266	8.5	779	15.2	3629
7.2	506	8.9	1162	10.7	1138	8.3	472	9.5	803	9.9	701	8.1	450	7.5	569
7.3	473	5.3	325	6.0	246	5.0	142	5.3	184	9.2	657	5.3	146	12.4	1830
11.8	1542	7.5	753	5.2	170	5.2	163	5.1	162	8.9	590	14.8	1395	7.4	816
6.8	465	6.0	443	7.1	546	6.8	331	6.8	391	7.5	434	9.3	881	9.8	1085
13.8	2151	7.1	753	7.8	310	6.9	329	7.1	486	7.1	555	6.8	421	15.8	3385
8.8	950	8.8	612	10.0	940	6.6	265	5.0	101	7.4	614	5.2	175	--	--
6.7	431	6.8	566	8.6	672	5.9	185	15.1	2791	6.8	325	7.4	582	--	--
7.5	956	5.4	411	9.1	1012	5.3	159	6.2	334	8.6	883	6.8	429	--	--
5.7	334	5.2	199	6.8	433	6.4	338	24.5	4797	7.1	526	8.1	477	--	--
7.6	436	6.1	255	11.8	2039	12.1	1531	9.3	795	12.1	1440	6.6	198	--	--
<u>Species group 3</u>															
6.5	535	5.4	373	5.5	281	6.5	284	6.7	540	8.4	167	5.3	320	9.4	1076
11.1	1023	12.6	1931	5.5	218	6.0	235	11.9	1435	6.9	151	8.8	1137	15.9	3606
13.7	2560	11.9	1438	5.0	307	19.2	3544	9.6	910	8.4	134	7.4	686	13.9	2999
9.6	941	6.7	685	6.3	462	7.2	524	10.8	965	8.8	261	15.6	3083	10.8	1445
6.6	527	6.6	548	6.5	415	12.2	1929	10.0	1053	7.0	508	7.8	704	6.7	409
9.0	1242	7.3	599	7.9	516	5.4	374	10.7	740	14.3	2877	10.9	1464	12.5	1361
13.0	2140	6.4	443	5.2	267	15.2	2736	8.5	510	9.2	1340	6.9	143	5.8	379
5.7	238	11.9	1464	7.5	645	12.7	1577	5.5	182	7.1	441	9.7	1063	6.5	405
5.9	327	6.0	367	7.1	581	7.8	676	24.7	6636	14.8	3287	12.8	1255	10.4	1069
5.1	197	13.6	2759	6.1	393	7.7	743	7.4	559	5.0	183	8.3	1053	8.6	825
22.6	6056	6.3	367	9.8	655	12.3	1710	7.4	449	5.8	398	14.3	2611	8.9	748
9.4	1576	7.9	715	5.6	422	7.0	541	8.8	1046	6.3	508	11.5	1531	10.1	1010
14.7	3744	5.3	308	6.8	446	10.7	1277	8.6	790	6.6	584	9.9	1110	14.5	2264
8.4	1106	6.5	511	7.3	488	7.6	546	9.2	678	7.3	521	10.3	841	--	--
9.7	1630	14.2	2819	6.9	404	15.8	2815	10.2	998	7.5	735	15.2	2233	--	--
5.9	423	12.4	1995	7.0	568	5.6	252	6.0	317	5.8	354	12.1	1847	--	--
6.0	321	10.2	1367	8.8	1116	14.3	2221	5.7	278	5.6	328	9.6	1241	--	--
7.0	1035	10.3	1090	5.1	217	5.7	278	8.8	1416	6.4	537	23.1	5804	--	--
6.0	568	6.5	354	8.6	773	10.8	1162	7.8	235	6.0	478	9.9	1398	--	--

for each individual sample tree $k = 1, 2, \dots$, 353 of Table 1. For example, for $k = 1$, the first tree of species 1, these variables are

$$u_1 = 557/(9.3)^2 = 6.4400509$$

$$x_1 = 1/(9.3)^2 = .011562030$$

$$x_2 = 1/9.3 = .10752688$$

$$x_3 = 1$$

The new variables u are arranged in a 353 by 1 vector $[U]$ and the new variables v in a 353 by 3 matrix $[V]$. We continue by calculating the matrices of cross products

$$[T] = [V]'[V] = \begin{bmatrix} .14868724 & .94619303 & 6.3324803 \\ .94619303 & 6.3324803 & 45.275493 \\ 6.3324803 & 45.275493 & 353 \end{bmatrix}$$

$$[P] = [V]'[U] = \begin{bmatrix} 58.746219 \\ 430.50963 \\ 3456.2479 \end{bmatrix}$$

the inverse of the matrix $[T]$

$$[T]^{-1} = \begin{bmatrix} 1085.6999 & -276.84567 & 16.031589 \\ -276.84567 & 72.496669 & -4.3320188 \\ 16.031589 & -4.3320188 & .27086279 \end{bmatrix}$$

and the three basic statistics

$$[b] = [T]^{-1}[P] = \begin{bmatrix} 5.1818118 \\ -25.653078 \\ 12.988357 \end{bmatrix}$$

$=$ estimate of $[\beta]$
 $S_{uu|v} = ([U]'[U] - [b]'[P])/(n-3) = 8.0278958$
 $=$ estimate of $\sigma_{uu|v}$, the conditional variance of the transformed variables $u = y/d^2$ given d

and

$$[S_{bb}] = S_{uu|v}[T]^{-1}$$

$$= \begin{bmatrix} 8715.8856 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

$=$ estimate of the covariance matrix $[\sigma_{bb}]$ of $[b]$.

Under the basic assumptions of the weighted least squares linear regression method, the estimators $[b]$, $S_{uu|v}$ and $[S_{bb}]$ are all unbiased.

If the conditional variance of y given d is required, one can estimate it by the formula

$$S_{yy|d} = d^4 S_{uu|v}$$

For example, if $d = 10$ inches, then

$$S_{yy|d} = (10)^4 (8.0278958) = 80278.958 \text{ square pounds}$$

It may be interesting to consider the null hypothesis $\beta_1 = 0$, that is, the regression function passes through the origin of the two axes. Because the test statistic t is equal to

$$t = b_1 / \sqrt{S_{b_1 b_1}} = 5.1818118 / \sqrt{8715.8856} = .06$$

a value which is not significant, the null hypothesis would normally be accepted. However, what happens at $d=0$ is irrelevant; we have no trees of diameter equal to zero and we should not extrapolate the application of the regression function anyway.

We can similarly test the null hypothesis that $\beta_2 = 0$. As the test statistic

$$t = b_2 / \sqrt{S_{b_2 b_2}} = -25.653078 / \sqrt{581.99570} = -1.06$$

is not significant, this null hypothesis would normally be accepted as well. However, the restricted parabolic model

$$\hat{y} = \beta_1 + \beta_3 d^2$$

includes all possible parabolas for which the minimum occurs at $d=0$ and the full parabolic model

$$\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$$

includes the set of all possible parabolas. As the least squares model selects the parabola of least squares, there seems no reason to work with the restricted when we can work with the full parabolic model.

Note that the two separate null hypotheses above are not equivalent to the null hypothesis $\beta_1 = \beta_2 = 0$. This is because, if one of two null hypotheses is accepted, and the corresponding term eliminated from the regression function, the other null hypothesis will not necessarily be accepted also.

Using the formula

$$[b]'[x] \pm t \sqrt{[x]'[S_{bb}][x]}$$

for the confidence interval, the formula

$$[b]'[x] \pm t \sqrt{d^4 S_{uu|v} + [x]'[S_{bb}][x]}$$

where

$$[x]' = [1 \quad d \quad d^2]$$

for $d = 5, 6, \dots, 30$

for the prediction interval of future values and $t=2$ for the 95 percent confidence level, we have calculated the biomass table and its 95 percent confidence and prediction limits shown in Table 2. The corresponding regression function and the 95 percent confidence and prediction limits are shown graphically in Figure 1.

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

- Cunia, T. On tree biomass tables and regressions: Some statistical comments. In: 1979 forest resource inventories workshop proceedings, vol. II, W. E. Frayer, (Ed.), Colorado State University, Fort Collins, CO; 1979a.
- Cunia, T. On sampling trees for biomass tables construction: some statistical comments. In: 1979 forest resource inventories workshop proceedings, vol. II, W. E. Frayer (Ed.), Colorado State University, Fort Collins, CO; 1979b.
- Cunia, T. Cluster sampling and tree biomass tables construction. In: Interdivisional Proceedings, 17th IUFRO World Congress, September 6-12, 1981, Kyoto, Japan; 1981.

Table 2 - The tree biomass table constructed by the weighted least squares method from sample data of Table 1 and its 95 percent confidence and prediction limits.

Diameter d	Lower 95 Percent Limits		Regression Estimates	Upper 95 Percent Limits	
	Prediction	Confidence		Confidence	Prediction
5	58	178	202	225	345
6	114	304	319	334	523
7	184	443	462	482	740
8	268	605	631	658	995
9	366	794	826	859	1287
10	479	1008	1048	1087	1616
11	607	1244	1295	1346	1982
12	749	1500	1568	1636	2387
13	905	1775	1869	1958	2829
14	1074	2070	2192	2313	3309
15	1258	2385	2543	2701	3828
16	1455	2719	2920	3121	4384
17	1666	3073	3323	3573	4979
18	1890	3446	3752	4057	5613
19	2128	3840	4207	4573	6285
20	2380	4254	4687	5121	6995
21	2644	4687	5194	5701	7744
22	2922	5141	5727	6314	8532
23	3214	5614	6286	6958	9358
24	3519	6108	6871	7633	10223
25	3837	6622	7481	8341	11126
26	4168	7155	8118	9081	12068
27	4513	7709	8781	9853	13049
28	4871	8283	9470	10656	14068
29	5242	8877	10184	11492	15126
30	5627	9491	10925	12359	16223

Cunia, T. On the error of biomass estimates in forest inventories; Part 1: Its major components. Faculty of Forestry Miscellaneous Publication Number 8 CESF 85-004), SUNY College of Environmental Science and Forestry, Syracuse, NY, 1985.

Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the workshop on "Tree Biomass regression functions and their contributions to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.

Cunia, T. An optimization model for subsampling trees for biomass measurement. In: Proceedings of the workshop on "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; 1986b.

Cunia, T. Evaluating errors of tree biomass regressions by simulation. In: Proceedings of the workshop on "Tree Biomass regression functions and their contributions to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY, 1986c.

Cunia, T. On the error of tree biomass regressions: trees selected by cluster sampling and double sampling. In: Proceedings of the workshop on "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; 1986d.

USE OF DUMMY VARIABLES TECHNIQUES IN THE ESTIMATION OF BIOMASS REGRESSIONS

Tibbeerius Cunia

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

If the sample elements can be subdivided into classes, one can use linear regression with dummy variables techniques to estimate individual class regressions and their error when their regression coefficients are interrelated. It is shown how to test null hypotheses about relationships among coefficients and it is illustrated how to apply these techniques to problems of (i) piecewise regressions, (ii) harmonization of biomass regressions and (iii) forcing additivity of biomass tables.

Introduction

In his paper, Cunia (1986a) reviews the method of weighted least squares linear regression as applied to the problem of estimating the regression of tree biomass y on several independent variables x defined in terms of diameter, height, etc. He assumed implicitly that all variables were quantitative, measured by an interval or ratio scale. There are cases, however, where the sample trees can be classified according to some criterion (species, forest type, site quality etc.) into several classes. Biomass regression functions may be required separately by classes or a single biomass regression may be desired for all classes, regression that includes the classification criterion as a set of independent variables x . In both cases it is useful to work with dummy variables techniques of the type described, among others, by Cunia (1973).

The objectives of the present paper are those of describing the general method of linear regressions with dummy variables and showing how this method has been applied to a variety of problems involving calculations of tree biomass regressions. More specifically we shall show how to use the standard, least or weighted least squares linear regression techniques and (i) estimate independent biomass regression functions for individual classes, (ii) test for similarity or identity the corresponding coefficients of various individual class regressions, (iii) calculate common estimators (and their error) of regression coefficients when the corresponding tests show that they are not significantly different and (iv) apply dummy variables to estimate piecewise linear regressions, harmonize biomass regressions and insure additivity of the regressions of several biomass components. In all cases we shall show how to calculate the vectors of regression coefficients $[b^i]$ of class i and the covariance matrices $[S_{bb}^{ij}]$ of $[b^i]$ and $[b^j]$; these

statistics are needed, see Cunia (1986b) to calculate the error of forest inventory estimates.

Estimation of Individual Class Regressions

Assume that the n elements (trees) of the sample are grouped into q mutually exclusive and collectively exhaustive classes and are measured for the dependent variable y and the independent variables $x_1, x_2, \dots, x_m$ which may be referred to as the vector of order m

$$[x]' = [x_1 \ x_2 \ \dots \ x_m]$$

Assume also that the subsample of n_i elements in class i satisfies the basic assumptions of the least squares linear regression of y on $[x]$, in particular that the regression is of the linear form

$$\hat{y}_i = \beta_{i1}x_1 + \beta_{i2}x_2 + \dots + \beta_{im}x_m = [\beta^i]'[x] \text{ for } i = 1, 2, \dots, q$$

and that the conditional variance of y given $[x]$ is homogeneous over all n elements of the sample

Let us define now

- (1) the dummy variables D_i , $i = 1, 2, \dots, q$
 $D_i = 1$ if the tree belongs to class i
 $= 0$ otherwise

- (2) the new variables x_{ij} (with the first subscript denoting the class number $i = 1, 2, \dots, q$ and the second subscript denoting the variable number $j = 1, 2, \dots, m$) as

$$x_{ij} = D_i x_j = x_j \text{ if the tree belongs to class } i \\ = 0 \text{ otherwise,}$$

- and (3) the giant size regression

$$\begin{aligned} \hat{y} &= \sum \beta_{ij} x_{ij} = [\beta]'[x] \\ &= \beta_{11}x_{11} + \beta_{12}x_{12} + \dots + \beta_{1m}x_{1m} \\ &\quad + \beta_{21}x_{21} + \beta_{22}x_{22} + \dots + \beta_{2m}x_{2m} \\ &\quad + \dots \\ &\quad + \beta_{q1}x_{q1} + \beta_{q2}x_{q2} + \dots + \beta_{qm}x_{qm} \end{aligned}$$

where

$$[\beta]' = [\beta_{11} \ \beta_{12} \ \dots \ \beta_{1m} \ \beta_{21} \ \dots \ \beta_{qm}] \\ = \left[[\beta^1]' \ [\beta^2]' \ \dots \ [\beta^q]' \right]$$

and a similar expression for $[x]$.

It is not necessary that the same independent variables be included in all regressions; by forcing coefficients β_{ij} to be equal to zero we can eliminate the variables x_j that are not desired in the regression of class i . It is also not necessary for a given variable x_j to have all the regression coefficients β_{ij} distinct; some classes may be defined as having equal regression coefficients for some variable x_j . Because the q individual class regressions are all included in

the giant size regression, one can (i) estimate all regressions in one single step and (ii) test null hypotheses about relationships among the coefficients of the various regressions, in particular hypotheses about equality of two or more coefficients β_{ij} (for given j).

Note that the giant size regression can be written in a long format form, showing in explicit terms the individual class regressions, as

$$\begin{aligned}\hat{y} &= \beta_{11}x_1 + \beta_{12}x_2 + \dots + \beta_{1m}x_m \\ &\quad \text{if the tree belongs to class 1} \\ &= \beta_{21}x_1 + \beta_{22}x_2 + \dots + \beta_{2m}x_m \\ &\quad \text{if the tree belongs to class 2} \\ &= \dots \\ &= \beta_{q1}x_1 + \beta_{q2}x_2 + \dots + \beta_{qm}x_m \\ &\quad \text{if the tree belongs to class } q\end{aligned}$$

where some β_{ij} may be equal to zero and not all β_{ij} are distinct.

A simple procedure to calculate the least squares estimators $[B]$ of $[\beta]$ and $[S_{BB}]$ of the covariance matrix $[\sigma_{BB}]$ of $[B]$ is that described by Cunia (1986a). We start by defining the matrices $[Y]$ and $[X]$ of sample data and the matrices $[T] = [X]'[X]$ and $[P] = [X]'[Y]$ of the sums of crossproducts. If $[T]^{-1}$ denotes the inverse of the matrix $[T]$, then

$$\begin{aligned}[B] &= [T]^{-1}[P] \\ S_{yy|x} &= ([Y]'[Y] - [B]'[P])/(n-qm) \\ \text{and} \quad [S_{BB}] &= S_{yy|x}[T]^{-1}\end{aligned}$$

Note that the number of degrees of freedom $(n-qm)$ of $S_{yy|x}$ is correct only when each individual class regression has m independent variables (including the intercept terms $x_{i1} = 1$). The vector $[B]$ can be written in terms of the subvectors of regression coefficients $[b^i]$ as

$$\begin{aligned}[B]' &= \begin{bmatrix} [b^1]' & [b^2]' & \dots & [b^q]' \end{bmatrix} \\ &= [b_{11} \quad b_{12} \quad \dots \quad b_{1m} \quad b_{21} \quad b_{22} \quad \dots \quad b_{qm}]\end{aligned}$$

and the matrix $[S_{BB}]$ can be expressed in terms of the covariance matrices $[S_{bb}^{ij}]$ of $[b^i]$ and $[b^j]$ as

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [S_{bb}^{12}] & \dots & [S_{bb}^{1q}] \\ [S_{bb}^{21}] & [S_{bb}^{22}] & \dots & [S_{bb}^{2q}] \\ \vdots & \vdots & & \vdots \\ [S_{bb}^{q1}] & [S_{bb}^{q2}] & \dots & [S_{bb}^{qq}] \end{bmatrix}$$

When the subsamples of various classes are statistically independent, and distinct regression functions are fit in each class, the submatrices

$[S_{bb}^{ij}]$ are all equal to zero, as $[b^i]$ and $[b^j]$ are statistically independent.

When the conditional variance of y given $[x]$ is not homogeneous, one should normally apply the weighted least squares method. In the case of biomass regressions, one can apply the procedure described, among others by Cunia (1986a) which assumes that the conditional variance of y given $[x]$ is proportional to a value a^2 known for all sample (and population) elements. Let us assume more specifically that

$$\sigma_{yy|x} = \sigma_{uu|v} a^2$$

where, for the trees with equal values $[x]$, we have the same function a^2 (independent of the class of trees) and $\sigma_{uu|v}$ is the same for all classes. Then, to calculate the weighted least squares estimates, we define first the transformed variables $u=y/a$, $v_{ij}=x_{ij}/a$ arranged in the matrices $[V]$ and $[U]$ of sample data. If $[T] = [V]'[V]$ and $[P] = [V]'[U]$, then

$$[B] = [T]^{-1}[P]$$

$$S_{uu|v} = ([U]'[U] - [B]'[P])/(n-qm)$$

$$\text{and } [S_{BB}] = S_{uu|v}[T]^{-1}$$

Ordinarily $\sigma_{uu|v}$ would change from class to class. This is not important, however, since the differences are relatively small and the weighted least squares method is not too sensitive to small departures from its assumption of homogeneity of conditional variance of the transformed variable u . But if the departure is felt to be large and one wishes to take it into account, the procedure is more complex and requires the following steps

Step 1 - Calculate each class regression separately. Let the estimate $S_{uu|v}$ of class i be denoted as $S_{uu|v}^i$.

Step 2 - Proceed with the weighted least squares method as usual, but define the new transformed variables as

$$u = y/a\sqrt{S_{uu|v}^i} \quad \text{and} \quad v_{ij} = x_{ij}/a\sqrt{S_{uu|v}^i}$$

It may be of interest to note that for this case,

$$[U]'[U] - [B]'[P] = n-qm$$

and thus,

$$S_{uu|v} = 1 \quad \text{and} \quad [S_{BB}] = [T]^{-1}$$

Let us apply the above procedures to a numerical example.

Example 1 - Consider the data from the 353 sample trees listed in Table 1 of a paper by Cunia (1986a). These trees were measured for total above ground tree biomass y (pounds of green weight) and diameter at breast height d (inches). They were also classified into three species groups (1 for pines, 2 for maples and 3 for all other species). There are 100, 107 and 146 trees in species groups 1, 2 and 3 respec-

tively. It is desired to calculate a different biomass regression function for each individual group. From past empirical experience it is known that within each species group, (i) the biomass regression function is of the parabolic form, $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$ and (ii) the conditional variance of y given d is proportional to d^4 . It has been verified that these assumptions are also satisfied by our sample data.

To estimate the three regression functions we shall use the following two procedures

Procedure 1 - Applying the weighted least squares method as described by Cunia (1986a) to each species group $i = 1, 2, 3$ separately we obtain the following statistics, where the super-script refers to the group number i .

- (1) The matrices $[T^i]$ and $[P^i]$ of the sums of crossproducts of the variables u and v_{ij}

$$[T^1] = \begin{bmatrix} .04333 & .27451 & 1.83052 \\ .27451 & 1.83052 & 13.00988 \\ 1.83052 & 13.00988 & 100 \end{bmatrix}$$

$$[T^2] = \begin{bmatrix} .04917 & .30862 & 2.02798 \\ .30862 & 2.02798 & 14.15100 \\ 2.02798 & 14.15100 & 107 \end{bmatrix}$$

$$[T^3] = \begin{bmatrix} .05618 & .36306 & 2.47398 \\ .36306 & 2.47398 & 18.11461 \\ 2.47398 & 18.11461 & 146 \end{bmatrix}$$

$$[P^1] = \begin{bmatrix} 14.32026 \\ 104.79302 \\ 836.39548 \end{bmatrix}, [P^2] = \begin{bmatrix} 18.33304 \\ 131.75597 \\ 1027.3029 \end{bmatrix}$$

$$[P^3] = \begin{bmatrix} 26.09292 \\ 193.96063 \\ 1592.5495 \end{bmatrix}$$

- (2) The inverses of the matrices $[T^i]$,

$$[T^1]^{-1} = \begin{bmatrix} 4231.9155 & -1115.5300 & 67.66323 \\ -1115.5300 & 301.30214 & -18.77909 \\ 67.66323 & -18.77909 & 1.21455 \end{bmatrix}$$

$$[T^2]^{-1} = \begin{bmatrix} 3343.8895 & -863.61829 & 50.83846 \\ -863.61829 & 229.43526 & -13.97511 \\ 50.83846 & -13.97511 & .89404 \end{bmatrix}$$

$$[T^3]^{-1} = \begin{bmatrix} 2776.4363 & -687.93798 & 38.30736 \\ -687.93798 & 174.87134 & -10.03959 \\ 38.30736 & -10.03959 & .60337 \end{bmatrix}$$

- (3) The estimates of the vectors of regression coefficients and their covariance matrices

$$[b^1]' = [b_{11} \quad b_{12} \quad b_{13}] = [295.60183 \quad -107.06967 \quad 16.882552]$$

$$[b^2]' = [b_{21} \quad b_{22} \quad b_{23}] = [-256.70604 \quad 40.050701 \quad 9.1695394]$$

$$[b^3]' = [b_{31} \quad b_{32} \quad b_{33}] = [18.800242 \quad -20.693393 \quad 13.156786]$$

$$s^1_{uu|v} = 3.1325346$$

$$s^2_{uu|v} = 7.7487808$$

$$s^3_{uu|v} = 9.0755518$$

$$[S^1_{bb}] = \begin{bmatrix} 13256.622 & -3494.4363 & 211.95740 \\ -3494.4363 & 943.83938 & -58.826156 \\ 211.95740 & -58.826156 & 3.8046237 \end{bmatrix}$$

$$[S^2_{bb}] = \begin{bmatrix} 25911.067 & -6691.9889 & 393.93612 \\ -6691.9889 & 1777.8435 & -108.29003 \\ 393.93612 & -108.29003 & 6.9277171 \end{bmatrix}$$

$$[S^3_{bb}] = \begin{bmatrix} 25197.692 & -6243.4168 & 347.66044 \\ -6243.4168 & 1587.0539 & -91.114812 \\ 347.66044 & -91.114812 & 5.4758847 \end{bmatrix}$$

Procedure 2 - Assuming that the same conditional variance function applies to the sample trees of all three species groups, that is, same function a^2 and $\sigma_{uu|v}$, we start by defining the dummy variables

$$D_1 = 1 \text{ if tree of group 1}$$

$$= 0 \text{ otherwise}$$

$$D_2 = 1 \text{ if tree of group 2}$$

$$= 0 \text{ otherwise}$$

$$D_3 = 1 \text{ if tree of group 3}$$

$$= 0 \text{ otherwise}$$

If the single subscripted variables x_j are defined as

$$x_1 = 1, x_2 = d, x_3 = d^2$$

we defined the double subscripted variables x_{ij} as

$$x_{i1} = D_i x_1 = 1 \text{ if tree of species } i$$

$$= 0 \text{ otherwise}$$

$$x_{i2} = D_i x_2 = d \text{ if tree of species } i$$

$$= 0 \text{ otherwise}$$

$$x_{i3} = D_i x_3 = d^2 \text{ if tree of species } i$$

$$= 0 \text{ otherwise}$$

for $i = 1, 2, 3$.

Because of the assumption that the conditional variance of y given d is proportional to d^4 (that is, $a = d^2$), we define the new variables u and v_{ij} as

$$u = y/d^2 \text{ and } v_{ij} = x_{ij}/d^2$$

Note that

$$v_{i1} = 1/d^2 \text{ if tree of species } i$$

$$= 0 \text{ otherwise}$$

$$v_{i2} = 1/d \text{ if tree of species } i$$

$$= 0 \text{ otherwise}$$

$$v_{i3} = 1 \text{ if tree of species } i$$

$$= 0 \text{ otherwise}$$

The matrices $[T]$ and $[P]$ of the sums of crossproducts take the form

$$[T] = \begin{bmatrix} [T^1] & [0] & [0] \\ [0] & [T^2] & [0] \\ [0] & [0] & [T^3] \end{bmatrix}$$

and

$$[P] = \begin{bmatrix} [P^1] \\ [P^2] \\ [P^3] \end{bmatrix}$$

where $[T^i]$ and $[P^i]$ are the matrices of the sums of crossproducts of species i obtained by Procedure 1 above. The inverse of $[T]$ takes the form

$$[T]^{-1} = \begin{bmatrix} [T^1]^{-1} & [0] & [0] \\ [0] & [T^2]^{-1} & [0] \\ [0] & [0] & [T^3]^{-1} \end{bmatrix}$$

and, using matrix multiplication by blocks,

$$[B] = [T]^{-1}[P] = \begin{bmatrix} [T^1]^{-1}[P^1] \\ [T^2]^{-1}[P^2] \\ [T^3]^{-1}[P^3] \end{bmatrix} = \begin{bmatrix} [b^1] \\ [b^2] \\ [b^3] \end{bmatrix}$$

This is verified when the calculations are performed with the sample data. Then, it is found that

$$[B]' = \begin{bmatrix} 295.60183 & -107.06967 & 16.882552 \\ -256.70604 & 40.050701 & 9.1695394 \\ 18.800242 & -20.693393 & 13.156786 \end{bmatrix}$$

$$S_{uu|v} = 6.9986424$$

and the covariance matrix of $[B]$ is

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [0] & [0] \\ [0] & [S_{bb}^{22}] & [0] \\ [0] & [0] & [S_{bb}^{33}] \end{bmatrix}$$

where

$$[S_{bb}^{11}] = \begin{bmatrix} 29617.663 & -7807.1955 & 473.55072 \\ -7807.1955 & 2108.7059 & -131.42815 \\ 473.55072 & -131.42815 & 8.50021 \end{bmatrix}$$

$$[S_{bb}^{22}] = \begin{bmatrix} 23402.687 & -6044.1556 & 355.80023 \\ -6044.1556 & 1605.73531 & -97.80676 \\ 355.80023 & -97.80676 & 6.25706 \end{bmatrix}$$

and

$$[S_{bb}^{33}] = \begin{bmatrix} 19431.285 & -4814.6319 & 268.09952 \\ -4814.6319 & 1223.8620 & -70.26349 \\ 268.09952 & -70.26349 & 4.22275 \end{bmatrix}$$

Note that although the vectors $[b^i]$ of regression coefficients are identical by the two procedures, their covariance matrices are not. This is because in Procedure 2 we assume that the conditional variance of u given $[v]$ is the same in all strata. As this is not the case, and $[S_{bb}] = S_{uu|v}[T]^{-1}$, the covariance matrices $[S_{bb}]$ are not the same, even though the inverse matrices $[T]^{-1}$ are identical.

It may be interesting to realize that we have the following relationships

$$S_{uu|v} = \frac{(n_1-3)S_{uu|v}^1 + (n_2-3)S_{uu|v}^2 + (n_3-3)S_{uu|v}^3}{(n_1-3) + (n_2-3) + (n_3-3)}$$

where (n_i-3) is the number of degrees of freedom of $S_{uu|v}^i$ and that

$$[S_{bb}^{ii}] = (S_{uu|v}/S_{uu|v}^i)[S_{bb}]$$

Because $[S_{bb}^{ii}]$ of Procedure 1 is not equal to $[S_{bb}^{ii}]$ of Procedure 2, a question arises as to which covariance matrix to use, when it is important to know the error of the biomass regressions or the error of forest inventory estimates based on these regressions. At first sight, it seems that the error is estimated more accurately by $[S_{bb}^{ii}]$. A look at the three sample values $S_{uu|v}^i$ shows that $S_{uu|v}^1=3.13$ of species group 1 (pine trees) is the smallest, followed by $S_{uu|v}^2=7.75$ of species group 2 (maple trees) and $S_{uu|v}^3=9.08$ of species group 3 (all remaining trees). This is not surprising since softwood trees (pines) are expected to be less variable than hardwood trees (maples), and a species group consisting of trees of a variety of species is expected to be the most heterogeneous.

Consequently, when it is important to know the error of the individual regression functions, one is advised to use $[S_{bb}^{ii}]$. This may be the case when individual biomass tables are to be (i) constructed with their confidence and prediction limits and (ii) applied to estimate the error of forest inventory estimates by species group. But if the interest lies with the calculation of biomass estimates for all species combined, it does not really matter which matrix is used; the error of these estimates will be approximately the same. However, because working with $[S_{bb}^{ii}]$ presents advantages not shared by $[S_{bb}^{ii}]$ we prefer using the estimates calculated by Procedure 2.

One of the main advantages of dummy variables techniques and giant size regressions (of Procedure 2) is that of using the entire set of all sample tree data to (i) test null hypotheses about the values taken on by the coefficients of some variable x_j in one or more individual class regressions (as, for example, whether these coefficients are equal to zero, have common values or are related in some specific way), (ii) estimate the values of the coefficients of some variable x_j , when based on a significance test they can be assumed to be in a given relationship and (iii) to estimate the covariance matrix $[S_{bb}^{ij}]$ when $[b^i]$ and $[b^j]$ have common coefficients or, in general, are not statistically independent.

Testing Null Hypotheses about Regression Coefficients

Cunia (1986a) describes the way to test the general null hypothesis that several regression coefficients are all equal to zero. The same test applies, albeit in a different form, to null hypotheses about linear combinations of regression coefficients.

To describe this new form of significance test let us assume that we deal with a given, unrestricted regression R_1 defined as

$$R_1: \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

The null hypothesis is expressed as the following $r < m$ linear combinations among the m coefficients

$$\begin{aligned} a_{11}\beta_1 + a_{12}\beta_2 + \dots + a_{1m}\beta_m &= [a_1]'[\beta] = 0 \\ a_{21}\beta_1 + a_{22}\beta_2 + \dots + a_{2m}\beta_m &= [a_2]'[\beta] = 0 \\ &\vdots \\ a_{r1}\beta_1 + a_{r2}\beta_2 + \dots + a_{rm}\beta_m &= [a_r]'[\beta] = 0 \end{aligned}$$

Assume that we can solve the r equations in m variables β_i for any set of r regression coefficients in terms of the remaining $(m-r)$. Substitute this solution in R_1 , rearrange the terms under the remaining $(m-r)$ regression coefficients and, by using a different notation for the coefficients and the corresponding new independent variables, define the restricted (under the null hypothesis) regression R_2 as

$$R_2: \hat{y} = \beta'_1 x'_1 + \beta'_2 x'_2 + \dots + \beta'_{m-r} x'_{m-r}$$

We can now test the null hypothesis with the test statistic

$$F = \frac{(CR_1SS - CR_2SS)/r}{U_1SS/(n-m)}$$

which has the F -distribution with r and $(n-m)$ degrees of freedom, where

$$\begin{aligned} CR_1SS &= [b_1]'[P_1] \\ &= \text{regression sum of squares of } R_1 \end{aligned}$$

$$\begin{aligned} CR_2SS &= [b_2]'[P_2] \\ &= \text{regression sum of squares of } R_2 \end{aligned}$$

$$\begin{aligned} U_1SS &= [Y]'[Y] - CR_1SS \\ &= \text{unexplained sum of squares of } R_1 \end{aligned}$$

and n = sample size.

For the case of weighted least squares method all of the sums of squares above refer to the transformed variables u and v

Let us now illustrate how to apply this test to the giant size regression function of the previous section.

Example 2 - Consider the sample data and the giant size regression of Example 1. The reader can verify that additional calculations yield

$$\begin{aligned} [U]'[U] &= \sum u^2 = \sum (y/d^2) = 36961.259 \\ CR_1SS &= [B]'[P] = 34553.726 \\ U_1SS &= 36961.259 - 34553.726 = 2407.533 \\ U_1SS/(n-m) &= 2407.533/(353-9) = 6.9986424 \end{aligned}$$

where R_1 is the giant size regression

$$\begin{aligned} R_1: \hat{y} &= \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13} \\ &+ \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{23}x_{23} \\ &+ \beta_{31}x_{31} + \beta_{32}x_{32} + \beta_{33}x_{33} \end{aligned}$$

We would like to test now the null hypothesis

$$\beta_{12} = \beta_{22} = \beta_{32} = \beta_2 \text{ (same common value), and}$$

$$\beta_{13} = \beta_{23} = \beta_{33} = \beta_3 \text{ (same common value)}$$

When this null hypothesis is true, the three biomass regression functions are "parallel"; that is, they differ only by their intercept. The null hypothesis can also be expressed as the set of four linear combinations

$$\beta_{12} - \beta_{22} = [a_1]'[\beta] = 0$$

$$\beta_{22} - \beta_{32} = [a_2]'[\beta] = 0$$

$$\beta_{13} - \beta_{23} = [a_3]'[\beta] = 0, \text{ and}$$

$$\beta_{23} - \beta_{33} = [a_4]'[\beta] = 0$$

where

$$[a_1]' = [0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0]$$

$$[a_2]' = [0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 0]$$

$$[a_3]' = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0]$$

and

$$[a_4]' = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad -1]$$

Solving for β_{22} , β_{32} , β_{23} and β_{33} in terms of β_{12} and β_{13} , which we shall now denote by β_2 and β_3 respectively, we can write

$$\begin{aligned} R_2: \hat{y} &= \beta_{11}x_{11} + \beta_2 x_{12} + \beta_3 x_{13} \\ &+ \beta_{21}x_{21} + \beta_2 x_{22} + \beta_3 x_{23} \\ &+ \beta_{31}x_{31} + \beta_2 x_{32} + \beta_3 x_{33} \\ &= \beta_{11}x_{11} + \beta_{21}x_{21} + \beta_{31}x_{31} + \beta_2 x_2 + \beta_3 x_3 \end{aligned}$$

where

$$x_{12} + x_{22} + x_{32} = x_2, \text{ and}$$

$$x_{13} + x_{23} + x_{33} = x_3$$

The statistics associated with the restricted regression R_2 , are the following

$$[T_2] = \begin{bmatrix} .04333 & 0 & 0 & .27451 & 1.83052 \\ 0 & .04917 & 0 & .30862 & 2.02798 \\ 0 & 0 & .05618 & .36306 & 2.47398 \\ .27451 & .30862 & .36306 & 6.33248 & 45.27549 \\ 1.83052 & 2.02798 & 2.47398 & 45.27549 & 353 \end{bmatrix}$$

$$[P_2]' = [14.3203 \quad 18.3330 \quad 26.09292 \quad 430.5096 \quad 3456.2479]$$

Note that there is a relationship between the matrices $[T_2]$ and $[P_2]$ of the restricted regression R_2 and the corresponding matrices $[T_1]$ and $[P_1]$ of the unrestricted regression R_1 (Procedure 2, Example 1). Some values are the same, as for example .04333, .04917, etc., others are the sums of three values, as for example, 6.33248 (row 4, column 4 of $[T_2]$) which is equal to the sum of the three values 1.83052, 2.02798 and 2.47398 (row 2, column 2) of the submatrices $[T^1]$, $[T^2]$ and $[T^3]$ respectively of $[T_1]$ of the unrestricted regression R_1 .

After calculating $[T_2]^{-1}$, we find

$$[B_2]' = \begin{bmatrix} -36.543351 & 17.250897 & 77.620963 \\ -29.919713 & 13.174939 \end{bmatrix}$$

$$CR_2SS = [B_2]'[P_2] = 34473.439$$

and,

$$[S_{bb}] = \begin{bmatrix} 7856.12 & 7688.97 & 7736.77 & -1977.43 & 114.489 \\ 7688.97 & 7832.29 & 7734.47 & -1977.83 & 114.599 \\ 7736.77 & 7734.47 & 7910.14 & -1987.70 & 114.949 \\ -1977.43 & -1977.83 & -1987.70 & 518.766 & -30.9890 \\ 114.489 & 114.599 & 114.949 & -30.9890 & 1.93720 \end{bmatrix}$$

The test statistic is

$$F = \frac{(CR_2SS - CR_2SS)/(4)}{U_1SS/(353-9)} = 2.868$$

with 4 and 344 degrees of freedom. As the critical F values for the .05 and .01 probability of rejection are 2.37 and 3.32 respectively, the usual decision would be to reject the null hypothesis. However, for the purpose of illustrating how to express the results as a 9 by 1 giant size vector [B] (that contains the three biomass regressions in an explicit form) we shall assume that the null hypothesis is accepted and we use the statistics of the restricted regression R_2 . We can write the estimate of the restricted regression as

$$\begin{aligned} R_2: \hat{y} &= b_{11} + b_2d + b_3d^2 \text{ for species group 1} \\ &= b_{21} + b_2d + b_3d^2 \text{ for species group 2} \\ &= b_{31} + b_2d + b_3d^2 \text{ for species group 3} \end{aligned}$$

This implies that the biomass regressions of the three species groups and the corresponding covariance matrices are

$$[b^1] = \begin{bmatrix} b_{11} \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} -36.543351 \\ -29.919713 \\ 13.174939 \end{bmatrix}$$

$$[b^2] = \begin{bmatrix} b_{21} \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 17.250897 \\ -29.919713 \\ 13.174939 \end{bmatrix}$$

$$[b^3] = \begin{bmatrix} b_{31} \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 77.620963 \\ -29.919713 \\ 13.174939 \end{bmatrix}$$

$$\begin{aligned} [S_{bb}^{11}] &= \begin{bmatrix} S_{b_{11}b_{11}} & S_{b_{11}b_2} & S_{b_{11}b_3} \\ S_{b_2b_{11}} & S_{b_2b_2} & S_{b_2b_3} \\ S_{b_3b_{11}} & S_{b_3b_2} & S_{b_3b_3} \end{bmatrix} \\ &= \begin{bmatrix} 7856.12 & -1977.43 & 114.489 \\ -1977.43 & 518.766 & -30.9890 \\ 114.489 & -30.9890 & 1.93720 \end{bmatrix} \end{aligned}$$

and similarly

$$[S_{bb}^{22}] = \begin{bmatrix} 7832.29 & -1977.83 & 114.599 \\ -1977.83 & 518.766 & -30.9890 \\ 114.599 & -30.9890 & 1.93720 \end{bmatrix}$$

and

$$[S_{bb}^{33}] = \begin{bmatrix} 7910.14 & -1987.70 & 114.949 \\ -1987.70 & 518.766 & -30.9890 \\ 114.949 & -30.9890 & 1.93720 \end{bmatrix}$$

Because $[b^i]$ and $[b^j]$ are not statistically independent (they have common values b_2 and b_3) we have $[S_{bb}^{ij}] \neq 0$. Then, the reader can verify that

$$\begin{aligned} [S_{bb}^{12}] &= \begin{bmatrix} S_{b_{11}b_{21}} & S_{b_{11}b_2} & S_{b_{11}b_3} \\ S_{b_{21}b_{11}} & S_{b_{21}b_2} & S_{b_{21}b_3} \\ S_{b_2b_{11}} & S_{b_2b_{21}} & S_{b_2b_3} \end{bmatrix} \\ &= \begin{bmatrix} 7688.97 & -1977.43 & 114.489 \\ -1977.83 & 518.766 & -30.9890 \\ 114.599 & -30.9890 & 1.93720 \end{bmatrix} \end{aligned}$$

and similarly

$$\begin{aligned} [S_{bb}^{13}] &= \begin{bmatrix} 7736.77 & -1977.43 & 114.489 \\ -1987.70 & 518.766 & -30.9890 \\ 114.949 & -30.9890 & 1.93720 \end{bmatrix} \\ [S_{bb}^{23}] &= \begin{bmatrix} 7734.47 & -1977.83 & 114.599 \\ -1987.70 & 518.766 & -30.9890 \\ 114.949 & -30.9890 & 1.93720 \end{bmatrix} \end{aligned}$$

and

$$[S_{bb}^{21}] = [S_{bb}^{12}]', [S_{bb}^{31}] = [S_{bb}^{13}]' \text{ and } [S_{bb}^{32}] = [S_{bb}^{23}]'.$$

Consequently, the giant size vector [B] containing the subvectors of regression coefficients of the three individual species groups is the 9 by 1 vector

$$[B]' = \begin{bmatrix} [b^1]' & [b^2]' & [b^3]' \end{bmatrix}$$

and its covariance matrix is estimated by the 9 by 9 matrix

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [S_{bb}^{12}] & [S_{bb}^{13}] \\ [S_{bb}^{21}] & [S_{bb}^{22}] & [S_{bb}^{23}] \\ [S_{bb}^{31}] & [S_{bb}^{32}] & [S_{bb}^{33}] \end{bmatrix}$$

The numerical values of these subvectors and submatrices are shown above.

Piecewise Linear Regressions

In the previous sections the sample elements were classified into groups by a criterion other than their values $[x]$. This is not necessary, as sometimes classes may be defined in terms of the variables $[x]$. Then, the resulting regressions will be of the piecewise form, that is regressions for which the function changes its form over the range of independent variables $[x]$.

For example, when estimating biomass regressions one may have difficulties finding a single mathematical function that would fit sufficiently well the data over the entire range of tree diameters. The fit may be poor for the small trees, or for the large trees or for both. In this case, one may be able to fit piecewise regression functions consisting of one branch for the small and a second branch for the large trees; or, if the relationship seems to be of an S-shaped form, to fit a parabolic branch for the left-hand side, a linear branch for the middle and a second parabolic branch for the right-hand side of the regression function.

Fitting different mathematical functions to different classes of tree diameter would generally result in regression functions that present points of discontinuity at the class border; and this is not desirable. Then, one can set conditions on the choice of the regression coefficients of the various branches and, thus, obtain piecewise continuous regression functions. However, the transition from one to the next branch may not be smooth, even though the regression function is continuous; and this may also be undesirable. Then, one can force the two branches intersecting at the border of the two classes to have equal derivatives. In other words, one can set additional conditions on the choice of the regression coefficients and force the piecewise regression function to be continuous and have a first derivative everywhere.

Let us now show how to use linear regression with dummy variables to fit piecewise linear functions consisting of two branches. The extension to more than two branches is straightforward. We shall start with the grouping of the sample trees into two classes according to their diameter; class 1 containing the small trees with diameter less than or equal to d_0 (say $d_0 = 5$ inches) and class 2 containing the remaining trees. We shall assume that we want to fit different parabolic regressions to the two classes, that is, we shall make the assumption that

$$\hat{y} = \beta_{11} + \beta_{12}d + \beta_{13}d^2 \quad \text{if } d \leq d_0$$

$$= \beta_{21} + \beta_{22}d + \beta_{23}d^2 \quad \text{if } d > d_0$$

This can be written as the giant size regression

$$R_1: y = \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13}$$

$$+ \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{23}x_{23}$$

where

$$x_{11} = 1 \quad \text{if } d \leq d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{12} = d \quad \text{if } d \leq d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{13} = d^2 \quad \text{if } d \leq d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{21} = 1 \quad \text{if } d > d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{22} = d \quad \text{if } d > d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{23} = d^2 \quad \text{if } d > d_0, \text{ or } 0 \text{ otherwise}$$

The two parabolic branches will not general-

ly intersect at the point $d=d_0$. To force them to intersect at that point we shall make the assumption that the estimates of the biomass of the trees of diameter $d=d_0$ by the two regression branches are equal, that is

$$\beta_{11} + \beta_{12}d_0 + \beta_{13}d_0^2 = \beta_{21} + \beta_{22}d_0 + \beta_{23}d_0^2$$

This can be written as

$$[a]'[\beta] = 0$$

where

$$[a]' = [1 \quad d_0 \quad d_0^2 \quad -1 \quad -d_0 \quad -d_0^2]$$

Solving for β_{21} in terms of the other β coefficients, substituting this solution for β_{21} of regression R_1 , rearranging the terms by regression coefficients and by redefining the variables x we obtain the regression R_2 as follows:

$$\beta_{21} = \beta_{11} + \beta_{12}d_0 + \beta_{13}d_0^2 - \beta_{22}d_0 - \beta_{23}d_0^2$$

$$\hat{y} = \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13}$$

$$+ (\beta_{11} + \beta_{12}d_0 + \beta_{13}d_0^2 - \beta_{22}d_0 - \beta_{23}d_0^2)x_{21}$$

$$+ \beta_{22}x_{22} + \beta_{23}x_{23}$$

$$= \beta_{11}(x_{11}+x_{21}) + \beta_{12}(x_{12}+d_0x_{21}) + \beta_{13}(x_{13}+d_0^2x_{21})$$

$$+ \beta_{22}(x_{22}-d_0x_{21}) + \beta_{23}(x_{23}-d_0^2x_{21})$$

and

$$R_2: \hat{y} = \beta_{11}x'_{11} + \beta_{12}x'_{12} + \beta_{13}x'_{13}$$

$$+ \beta_{22}x'_{22} + \beta_{23}x'_{23}$$

where

$$x'_{11} = x_{11} + x_{21} = 1$$

$$x'_{12} = x_{12} + d_0x_{21} = d \quad \text{if tree diameter } \leq d_0$$

$$= d_0 \quad \text{otherwise}$$

$$x'_{13} = x_{13} + d_0^2x_{21} = d^2 \quad \text{if tree diameter } \leq d_0$$

$$= d_0^2 \quad \text{otherwise}$$

$$x'_{22} = x_{22} - d_0x_{21} = 0 \quad \text{if tree diameter } \leq d_0$$

$$= (d-d_0) \quad \text{otherwise}$$

$$x'_{23} = x_{23} - d_0^2x_{21} = 0 \quad \text{if tree diameter } \leq d_0$$

$$= (d^2-d_0^2) \quad \text{otherwise}$$

Of course, R_2 can also be written in terms of the individual class regressions as

$$\hat{y} = \beta_{11} + \beta_{12}d + \beta_{13}d^2 \quad \text{if tree diameter } \leq d_0$$

$$= \beta_{21} + \beta_{22}d + \beta_{23}d^2 \quad \text{otherwise}$$

with $\beta_{21} = \beta_{11} + \beta_{12}d_0 + \beta_{13}d_0^2 - \beta_{22}d_0 - \beta_{23}d_0^2$

Note that R_1 presents a discontinuity (a jump in the estimate of the biomass) at $d=d_0$ and that R_2 is a continuous function everywhere. If one wishes to test the null hypothesis that the size of the jump of regression R_1 at $d=d_0$ is equal to zero, he can use the F-statistic

$$F = (n-6)(CR_1SS - CR_2SS)/U_1SS$$

with 1 and n-6 degrees of freedom, where CRSS and USS are the regression and unexplained sums of squares. Of course, this test does not make sense in our case since we do not expect a jump. Consequently, the form of the regression function of tree biomass on diameter is that of R_2 .

Although continuous, the regression R_2 may present a sharp change as we move from one to the other branch of the regression at $d=d_0$. This may be undesirable. To obtain a smooth transition point we shall assume a second relationship among the regression coefficients; that the derivatives of the two parabolic branches are equal at $d=d_0$. More formally we make the assumption that in regression R_2 we have equality of the two derivatives at $d=d_0$, that is

$$\beta_{12} + 2d_0 \beta_{13} = \beta_{22} + 2d_0 \beta_{23}$$

that is $[c]'[\beta] = 0$

$$\text{where } [c]' = [0 \ 1 \ 2d_0 \ -1 \ -2d_0]$$

Solving for β_{22} , substituting this solution in R_2 , rearranging the terms and redefining the new variables, we obtain the new regression R_3 as follows

$$\begin{aligned} \beta_{22} &= \beta_{12} + 2d_0 \beta_{13} - 2d_0 \beta_{23} \\ \hat{y} &= \beta_{11}x'_{11} + \beta_{12}x'_{12} + \beta_{13}x'_{13} \\ &\quad + (\beta_{12} + 2d_0 \beta_{13} - 2d_0 \beta_{23})x'_{22} + \beta_{23}x'_{23} \\ &= \beta_{11}x'_{11} + \beta_{12}(x'_{12} + x'_{22}) \\ &\quad + \beta_{13}(x'_{13} + 2d_0 x'_{22}) + \beta_{23}(x'_{23} - 2d_0 x'_{22}) \end{aligned}$$

that is,

$$R_3: \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

where

$$x_1 = 1 \text{ for all trees}$$

$$x_2 = d \text{ for all trees}$$

$$x_3 = d^2 \text{ if tree diameter } \leq d_0$$

$$= d_0(2d - d_0) \text{ otherwise}$$

$$x_4 = 0 \text{ if tree diameter } \leq d_0$$

$$= (d - d_0)^2 \text{ otherwise}$$

In terms of the individual class regressions, R_3 can be written as

$$\begin{aligned} R_3: \hat{y} &= \beta_{11} + \beta_{12}d + \beta_{13}d^2 \text{ if tree diameter } \leq d_0 \\ &= \beta_{21} + \beta_{22}d + \beta_{23}d^2 \text{ if tree diameter } > d_0 \end{aligned}$$

where

$$\beta_{11} = \beta_1, \beta_{12} = \beta_2, \beta_{13} = \beta_3,$$

$$\beta_{21} = \beta_1 + d_0^2(\beta_4 - \beta_3)$$

$$\beta_{22} = \beta_2 + 2d_0(\beta_3 - \beta_4) \text{ and } \beta_{23} = \beta_4$$

The individual class regression coefficients can also be written, for $i=1,2$ and $j=1,2,3$ as

$$\text{where } \beta_{ij} = [a_{ij}]'[\beta]$$

$$[\beta] = [\beta_1 \ \beta_2 \ \beta_3 \ \beta_4]$$

$$[a_{11}]' = [1 \ 0 \ 0 \ 0]$$

$$[a_{12}]' = [0 \ 1 \ 0 \ 0]$$

$$[a_{13}]' = [0 \ 0 \ 1 \ 0]$$

$$[a_{21}]' = [1 \ 0 \ -d_0^2 \ d_0^2]$$

$$[a_{22}]' = [0 \ 1 \ 2d_0 \ -2d_0]$$

$$[a_{23}]' = [0 \ 0 \ 0 \ 1]$$

The procedure to estimate the regression R_3 is to apply the least squares method to the sample data y, x_1, x_2, x_3 and x_4 (or the weighted least squares method to the transformed sample data u, v_1, v_2, v_3 and v_4). More specifically, we define the matrices of sample data $[Y]$ and $[X]$ (or $[U]$ and $[V]$ in the case of weighted least squares) and we calculate the matrices of the sums of crossproducts $[T]$ and $[P]$, the vector of regression coefficients $[b]$ and the covariance matrix $[S_{bb}]$ of $[b]$ by the usual formulae. This yields the four regression coefficients b_1, b_2, b_3 and b_4 and the estimates of all their variances and covariances.

It is necessary sometimes to show the individual class regressions in explicit form, that is, to define the giant size vector $[B]$ of regression coefficients as

$$\begin{aligned} [B]' &= \begin{bmatrix} [b^1]' & [b^2]' \end{bmatrix} \\ &= [b_{11} \ b_{12} \ b_{13} \ b_{21} \ b_{22} \ b_{23}] \end{aligned}$$

where $b_{ij} = [a_{ij}]'[b]$ were defined above. To calculate the variances and covariances of b_{ij} is much more complex; for each b_{ij} and b_{hk} one should use the formula

$$\text{Cov}(b_{ij}, b_{hk}) = [a_{ij}]'[S_{bb}][a_{hk}]$$

For example, the reader can verify that

$$\text{Cov}(b_{11}, b_{11}) = [a_{11}]'[S_{bb}][a_{11}] = S_{b_{11}b_{11}}$$

$$\text{Cov}(b_{11}, b_{12}) = [a_{11}]'[S_{bb}][a_{12}] = S_{b_{11}b_{12}}$$

$$\begin{aligned} \text{Cov}(b_{11}, b_{21}) &= [a_{11}]'[S_{bb}][a_{21}] \\ &= S_{b_{11}b_{21}} - d_0^2 S_{b_{11}b_{23}} + d_0^2 S_{b_{13}b_{21}} \end{aligned}$$

etc.

It may be a good idea, at this point to verify that our regression R_3 is continuous and that the two branches have equal derivatives at $d=d_0$. Then,

(1) R_3 is continuous at $d=d_0$ since

$$\begin{aligned} &\beta_{21} + \beta_{22}d_0 + \beta_{23}d_0^2 \\ &= \beta_1 - d_0^2\beta_3 + d_0^2\beta_4 + d_0\beta_2 + 2d_0^2\beta_3 - 2d_0^2\beta_4 + d_0^2\beta_4 \end{aligned}$$

$$= \beta_1 + d_0 \beta_2 + d_0^2 \beta_3$$

$$= \beta_{11} + \beta_{12} d_0 + \beta_{13} d_0^2$$

and

(2) R_3 has a derivative at $d=d_0$ since derivative of right-hand side branch =

$$= \beta_{22} + 2d_0 \beta_{23}$$

$$= \beta_2 + 2d_0 \beta_3 - 2d_0 \beta_4 + 2\beta_4 d_0$$

$$= \beta_2 + 2d_0 \beta_3 = \beta_{12} + 2d_0 \beta_{13}$$

= derivative of left-hand side branch

Note that, to apply this procedure one needs to know the value of d_0 . Ordinarily this is not the case. However, if one knows the approximate region (range of diameter values) where the regression function changes its form, one can use trial and error and determine that value of d_0 for which the sum of squared residuals is approximately minimized.

Sometimes one may have doubts about the curvilinearity of one or the other branch. To test the null hypothesis that $\beta_{13} = \beta_3 = 0$ one can use the test statistic $t = b_3 / \sqrt{S_{b_3 b_3}}$ which has the t-distribution with $(n-4)$ degrees of freedom. A similar test can be devised for the null hypothesis $\beta_{23} = \beta_4 = 0$. To test the null hypothesis that simultaneously $\beta_{13} = \beta_{23} = 0$, that is $\beta_3 = \beta_4 = 0$, one can use the F distribution; the unrestricted regression is R_3 and the restricted regression is the straight line

$$R_4: \hat{y} = \beta_1 + \beta_2 d$$

Another null hypothesis of interest may be that of two identical parabolic branches, that is,

$$\beta_{11} = \beta_{21}, \beta_{12} = \beta_{22} \text{ and } \beta_{13} = \beta_{23}$$

A first test starts with the giant size regression R_1 defined as the unrestricted regression and regression R_5

$$R_5: \hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$$

as the restricted one. The test statistic is

$$F = (n-6)(CR_1SS - CR_5SS) / (3)(U_1SS)$$

with 3 and $(n-6)$ degrees of freedom. A second test starts with regression R_3 defined as the unrestricted regression and R_5 as the restricted one. Note that in this approach the three equalities $\beta_{11} = \beta_{21}$, $\beta_{12} = \beta_{22}$ and $\beta_{13} = \beta_{23}$ reduce to the same equality $\beta_3 = \beta_4$. This is because $\beta_{11} = \beta_{21}$ can be written as

$$\beta_1 = \beta_1 + d_0(\beta_4 - \beta_3)$$

and this implies that $\beta_4 - \beta_3 = 0$. Similarly, for $\beta_{12} = \beta_{22}$

$$\beta_2 = \beta_2 + 2d_0(\beta_3 - \beta_4)$$

implies that $\beta_3 = \beta_4$. Finally, $\beta_{13} = \beta_{23}$ is a different way of expressing the same equality $\beta_3 = \beta_4$.

The restricted regression is R_3 because

$$\hat{y} = \beta_1 + \beta_2 d + \beta_3 (d^2 + 0) \text{ for tree diameters } \leq d_0$$

$$= \beta_1 + \beta_2 d + \beta_3 (0 + d^2) \text{ for tree diameters } > d_0$$

For more on piecewise regressions and an illustrative example, the reader is referred to Cunia (1973).

Harmonizing Biomass Regressions of Nested Components

A problem that arises quite often when biomass regression functions are estimated for various tree components, is that the estimated regressions may not behave the way they should with respect to each other. For example, if one tree component is a part of another component, their regressions should not intersect; or if they do, they should intersect outside the applicable range of tree diameter.

Jacobs and Cunia (1980) and Cunia and Briggs (1985a) have illustrated the application of linear regression with dummy variables techniques to the problem of harmonization of a set of regression functions estimating the biomass of the tree bole up to various top diameters. The regressions were forced to be similar in shape, the spacing between successive regressions had to behave rationally over the range of tree diameters of interest and the regressions were not allowed to cross over. It is the objective of the present section to illustrate the application of this approach to the harmonization of a specific set of nested components. In the present context, we shall define nested components as a set of successive components such that the biomass of one is smaller than, or at most equal to the biomass of the component preceding it in the set.

To better illustrate the technique let us consider the specific problem of estimating the biomass regression functions for the following six components; component 1 = entire tree bole, component 2 = bole up to 10 cm of top diameter, component 3 = bole up to 15 cm of top diameter, component 4 = bole up to 20 cm of top diameter, component 5 = bole up to 25 cm of top diameter and component 6 = bole up to 30 cm of top diameter. We have a sample of trees where each tree is measured for the biomass of at least one and as many as all six bole components.

An overall set of regression functions are desired, one function for each component so that (i) no part of a component may have more biomass than the component itself, (ii) the spacing between the regressions of the successive components should follow a reasonable pattern, (iii) the biomass estimates of the set should be increasing functions of tree diameter and decreasing functions of top diameter and (iv) no biomass estimate should be allowed to be negative within the applicable range of tree breast and top diameter.

To construct a regression model to fit the set of six regressions, one for each component, we shall start with a decision about the form of the regression function. Past experience has shown that (i) the regression function of the bole biomass y (of some component) on tree diameter at breast height d is of the parabolic form $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$ and (ii) the conditional variance of y given d is proportional to d^4 . Let us assume here that an analysis of the sample data shows that these assumptions are sufficiently well satisfied and, thus, they can be used as the assumptions of our basic regression model.

We shall now define the six dummy variables

$$D_i = 1 \text{ if component } i, \\ = 0 \text{ otherwise}$$

for $i=1, 2, \dots, 6$ and the doubly subscripted variables

$$x_{ij} = D_i x_j \text{ for all } i=1,2,\dots,6 \text{ and } j=1,2,3$$

where $x_1=1$, $x_2=d$ and $x_3=d^2$. The giant size regression can now be written as

$$R_1: \hat{y} = \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13} \\ + \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{23}x_{23} \\ + \dots \\ + \beta_{61}x_{61} + \beta_{62}x_{62} + \beta_{63}x_{63}$$

Note that, the giant size regression can also be written as a set of six individual component regressions

$$y = \beta_{11} + \beta_{12}d + \beta_{13}d^2 \text{ for component 1} \\ = \beta_{21} + \beta_{22}d + \beta_{23}d^2 \text{ for component 2} \\ \vdots \\ = \beta_{61} + \beta_{62}d + \beta_{63}d^2 \text{ for component 6}$$

Because all regression coefficients are now contained in the same giant size regression, we can force them to enter into specific relationships. For our specific problem, there are two approximate relationships that seem reasonable. These relationships can be expressed as null hypotheses and, thus, can be tested for significance.

Relationship 1: The difference between the biomass of two components of the same tree is independent of tree size. Expressed otherwise, the six regression functions are "parallel", that is, the "distance" (on the vertical direction) between two regression functions is the same over the entire range of tree diameters.

This relationship can be justified by the following arguments. The difference between two bole components of the same tree is a log with end diameters equal to the top diameters of the given two components. Of course, we assume here

implicitly that both components can be defined on the tree. It is also reasonable to expect the shape of the upper part of the tree bole to be approximately the same for the trees of all sizes. This implies that the length and the end diameter (and, thus, the biomass) of this log are approximately the same for the trees of all sizes. Of course, there are differences between individual trees of the same or of a different size. But on the average we shall expect to have, for a component $i < j = 1, 2, \dots, 6$,

$$\hat{y}_i - \hat{y}_j = (\beta_{i1} + \beta_{i2}d + \beta_{i3}d^2) - (\beta_{j1} + \beta_{j2}d + \beta_{j3}d^2) \\ = \text{a value which is independent of the tree diameter } d.$$

This implies that $\beta_{i2} = \beta_{j2}$, $\beta_{i3} = \beta_{j3}$ and the difference $(\beta_{i1} - \beta_{j1})$ represents the expected biomass of the average log expressing the difference between the two bole components i and j .

The null hypothesis that expresses this relationship in the form suitable for testing is

$$\beta_{12} = \beta_{22} = \dots = \beta_{62} \text{ and}$$

$$\beta_{13} = \beta_{23} = \dots = \beta_{63}$$

To test this null hypothesis, we apply the procedure outlined in a previous section. The giant size regression is the unrestricted regression R_1 and the restricted regression is

$$R_2: \hat{y} = \beta_{11}x_{11} + \beta_2 d + \beta_3 d^2 \\ + \beta_{21}x_{21} + \beta_{31}x_{31} + \dots + \beta_{61}x_{61}$$

Let us assume that this null hypothesis has been accepted and the new regression function is of the form R_2 . Then, we may continue with the testing of another relationship that seems also reasonable, namely

Relationship 2: The spacing between the regressions of successive bole components (which, under relationship 1 are "parallel") varies in a "quadratic" fashion. The spacing may be defined as the difference between the heights (the intercepts) of the regressions. If we assume that the intercepts β_{i1} are of the parabolic form

$$\beta_{i1} = \beta_1 + \beta_4 z_i + \beta_5 z_i^2$$

where z_i = top diameter = 0, 10, 15, ..., 30, the height of successive regressions would change in a quadratic fashion. This relationship can be justified by the following arguments.

Consider a large tree that has all six components and section the tree at the points of top diameter equal to 30, 25, 20, 15 and 10 cm. This yields six sections or logs, which starting from the top of the tree will be denoted here as log 1, 2, ..., 6. Note that bole component 1 contains all six logs, bole component 2 contains only the logs 2, 3, ..., 6, etc. and that each log (except log 6) represents the difference between two successive bole components. The volume (and, thus, the biomass) of each consecu-

tive log varies with the length (which may be a linear function of the top diameter) and the square of the log diameter (which goes for consecutive logs from 0 to 10, 15, ..., 30 cm).

The null hypothesis associated with this relationship is that, for some fixed parameters β_1 , β_4 , and β_5 , we have the relationships

$$\begin{aligned}\beta_{11} &= \beta_1 \\ \beta_{21} &= \beta_1 + 10\beta_4 + 100\beta_5 \\ \beta_{31} &= \beta_1 + 15\beta_4 + 225\beta_5 \\ &\vdots \\ \beta_{61} &= \beta_1 + 30\beta_4 + 900\beta_5\end{aligned}$$

or, in terms of the new variable $z_i = 0, 10, 15, \dots, 30$,

$$\beta_{i1} = \beta_1 + \beta_4 z_i + \beta_5 z_i^2$$

The restricted regression for this null hypothesis is

$$R_3: \hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2 + \beta_4 z + \beta_5 z^2$$

and the unrestricted regression is R_2 above.

For a numerical application of this technique, the reader is referred to Jacobs and Cunia (1980) who use the weighted least squares method to estimate the regressions and test the null hypotheses. Because the biomass of the various components of the same tree are not statistically independent, Cunia and Briggs (1985) use the generalized least squares method to estimate the same regressions and test the same null hypotheses.

Forcing Additivity of Biomass Regressions

If a tree component is subdivided into several mutually exclusive and collectively exhaustive subcomponents, it is usually desirable for the regression estimate of the biomass of the component to be equal to the sum of the regression estimates of the biomass of the subcomponents. For example, if we assume that proper definitions of the components are given, the estimate of the bole biomass should be equal to the sum of the estimates of the unmerchantable top biomass and the biomass of the merchantable bole.

To define the problem in more formal terms, let us assume that (i) the tree is subdivided into (s-1) mutually exclusive and collectively exhaustive components, (ii) the biomass of the component $i = 1, 2, \dots, (s-1)$ is denoted by y_i and the total biomass of the tree by y_s and (iii) for some independent variables $x_1, x_2, \dots, x_m$, the regression function of the component $i = 1, 2, \dots, s$ is denoted by

$$\hat{y}_i = r_i(x_1, x_2, \dots, x_m)$$

Then, it can be shown that (i) in the absence of

measurement error, the value y_s of a particular tree is equal to the sum of the values $y_1, y_2, \dots, y_{s-1}$ and (ii) the true regression value of the total y_s is equal to the sum of the true regression values $\hat{y}_i, i = 1, 2, \dots, (s-1)$.

Consider now a representative sample of n trees and its n sets of sample values $y_1, y_2, \dots, y_s$. It is then known that, if the regression functions of the component parts and that of the total are estimated separately, the resulting biomass regressions are not necessarily additive. And this is true in spite of the fact that, for every sample tree, we have $y_1 + y_2 + \dots + y_{s-1} = y_s$. To insure additivity is a problem considered, among others, by Kozak (1970), Cunia and Briggs (1984, 1985b), Chiyenda and Kozak (1984) and Reed and Green (1985).

One can recognize three specific methods to insure additivity.

Method 1 - The "best" (from a statistical point of view) regression function is calculated for each component $y_i, i = 1, 2, \dots, s-1$ separately, by whatever procedure is thought best. The regression function of the total y_s is then defined as the sum of the regression functions of the components, that is

$$\hat{y}_s = \hat{y}_1 + \hat{y}_2 + \dots + \hat{y}_{s-1}$$

The method can be used with linear, non linear or combinations of linear or non linear functions; different independent variables x can be used in different regressions; it is not necessary that all components be measured on all sample trees (that is, missing data for some components of some trees presents no problem); and it calculates the "best" regression for each component. On the other hand, the estimate of regression of the total is not necessarily the "best" and what is an important drawback sometimes, the method does not provide means for the estimation of the error of the total regression; the regressions of various components are not statistically independent and the error of their sum is practically impossible to estimate.

Method 2 - The additivity is obtained by using the same independent variables $x_1, x_2, \dots, x_m$ in the linear regression function of all components. If the weighted least squares method is used, the same weights must also be used in all regressions. The method has the main advantage that it is simple and easy to understand. On the other hand the regression functions of some components may not be the "best" in the statistical sense; regression coefficients may be estimated even though they may not be significantly different from zero; the same weights must be used with the weighted least squares method even when different weights may be better for different regressions, etc. Furthermore, if some data for some components are missing for some trees, the method fails to insure additivity; to obtain this additivity, one must ignore all information from the trees that have some data missing.

Method 3 - The additivity is obtained by

using linear regression techniques with dummy variables. More specifically we assume that the regression function of the biomass y on $x_1, x_2, \dots, x_m$ within a component i is of the linear form

$$\hat{y}_i = \beta_{i1}x_1 + \beta_{i2}x_2 + \dots + \beta_{im}x_m,$$

for $i = 1, 2, \dots, s$

These s regression functions are expressed together as the giant size regression

$$\begin{aligned}\hat{y} = & \beta_{11}x_{11} + \beta_{12}x_{12} + \dots + \beta_{1m}x_{1m} \\ & + \beta_{21}x_{21} + \beta_{22}x_{22} + \dots + \beta_{2m}x_{2m} \\ & + \dots \\ & + \beta_{s1}x_{s1} + \beta_{s2}x_{s2} + \dots + \beta_{sm}x_{sm}\end{aligned}$$

subject to the additivity constraints

$$\begin{aligned}\beta_{11} + \beta_{21} + \dots + \beta_{s-1,1} &= \beta_{s1} \\ \beta_{12} + \beta_{22} + \dots + \beta_{s-1,2} &= \beta_{s2} \\ &\vdots \\ \beta_{1m} + \beta_{2m} + \dots + \beta_{s-1,m} &= \beta_{sm}\end{aligned}$$

In this regression,

$$\begin{aligned}x_{ij} &= x_j \text{ if component } i \\ &= 0 \text{ otherwise}\end{aligned}$$

for all $i = 1, 2, \dots, s$ and $j = 1, 2, \dots, m$

It is not necessary for the regression function of each component to contain the same independent variables as all others; regression coefficients β_{ij} may be made equal to zero as desired; different sets of weights (in the weighted least squares method) may also be used for different regressions; missing biomass measurements for some components of some trees present no problem; the set of all regressions is "best" even though each individual regression is not necessarily the best; it is possible to measure the error of each regression separately using the information from data of all components. The fact that the biomass measurements of the various components of a given sample tree are not independent presents no problem as the generalized least squares method will take this dependence into account. The main drawback of the method is the fact that it is not well known and is not as easy to understand and apply as the other two methods.

Acknowledgements

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

Literature Cited

- Chiyenda, S. S.; Kozak, A. Additivity of component biomass regression equations when the underlying model is linear. *Canadian Journal of Forest Research*, Vol. 14:441-446; 1984.
- Cunia, T. Dummy variables and some of their uses in regression analysis. In: *Proceedings of the June 1973 meeting of IUFRO Subject Group S4.02, Nancy-France*, Vol. 1, T. Cunia, K. Kuusela and A. J. Nash, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1973.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: *Proceedings of the workshop on "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; 1986a.
- Cunia, T. Error of forest inventory estimates: its main components. In: *Proceedings of the workshop on "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; 1986b.
- Cunia, T.; Briggs, R. D. Forcing additivity of biomass tables: some empirical results. *Canadian Journal of Forest Research*, Vol. 14:376-384; 1984.
- Cunia, T.; Briggs, R. D. Harmonizing biomass tables by generalized least squares. *Canadian Journal of Forest Research*, Vol. 15:331-340; 1985a.
- Cunia, T.; Briggs, R. D. Forcing additivity of biomass tables: use of the generalized least squares method. *Canadian Journal of Forest Research*, Vol. 15:23-28; 1985b.
- Jacobs, M. W.; Cunia, T. Use of dummy variables to harmonize tree biomass tables. *Canadian Journal of Forest Research*, Vol. 10:483-490; 1980.
- Kozak, A. Methods for ensuring additivity of biomass components by regression analysis. *Forestry Chronicle*, Vol. 46:402-404; 1970.
- Reed, D. D.; Green, E. J. A method of forcing additivity of biomass tables when using a non-linear model. *Canadian Journal of Forest Research*, Vol. 15:1184-1187; 1985.

ON THE ERROR OF TREE BIOMASS REGRESSIONS: TREES
SELECTED BY CLUSTER SAMPLING AND DOUBLE SAMPLING

Tiberius Cunia

Professor of Statistics and Operations Research,
SUNY College of Environmental Science and Forestry,
Syracuse, NY, 13210

The least squares method used in the estimation of biomass regression functions requires that the sample tree values be statistically uncorrelated. This requirement is normally satisfied when the trees are selected by simple random sampling. However, this is seldom, if ever the case in real life; the trees are more efficiently selected in clusters and by methods other than simple random sampling. As the computer packages used in the least squares estimation of the biomass regression functions assume that the tree biomass values are uncorrelated, valid questions are raised about the validity of the estimators $[b]$ of the vector of regression coefficients and $[S_{bb}]$ of the covariance matrix of $[b]$. To insure the validity of these estimators one would normally have to modify the least squares techniques so as to take into account the effect of the method by which the sample trees were selected. Two such modifications are presented; one that can be applied when the trees are selected by cluster random sampling, the other, when the trees are selected by a two-phase or double sampling procedure.

Introduction

In his paper, Cunia (1986a) describes an approach to combine the error of biomass regression functions with the error from sample plots when forest biomass estimates are calculated and their error estimated. This approach requires that (i) the true biomass regression function be of the linear form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

where y is the biomass (of some tree component) and $[x]$ is a vector of known attributes other than biomass and (ii) valid estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ are given. Note that we have used $[]$ and $[]'$ to denote vectors or matrices and their transposes respectively.

The standard least squares method of linear regression as needed here is described by Cunia (1986b). He (i) defines the basic matrices $[X]$ and $[Y]$ of the values $[x]$ and y of the n sample elements and (ii) calculates the estimators $[b]$ of the vector $[\beta]$ of regression estimators and $[S_{bb}]$ of the covariance matrix of $[b]$ by the formulae

$$[b] = [T]^{-1}[P] \text{ and } [S_{bb}] = S_{yy|x}[T]^{-1}$$

where

$$[T] = [X]'[X]$$

$$[P] = [X]'[Y]$$

$[T]^{-1}$ is the inverse of $[T]$, and

$$S_{yy|x} = ([Y]'[Y] - [b]'[P])/(n-m)$$

= estimator of the conditional variance of y given $[x]$

In the same paper, Cunia (1986b) describes the changes to make in the procedure when the least is replaced by the weighted least squares method. Assuming that the conditional variance of y given $[x]$ is proportional to the known value a^2 , we start by defining the new, transformed variables $u = y/a$ and $[v] = [x]/a$. These new variables are arranged in the new matrices $[V]$ and $[U]$ of sample data and the estimators $[b]$ and $[S_{bb}]$ are defined by the formulae

$$[b] = [T]^{-1}[P] \text{ and}$$

$$[S_{bb}] = S_{uu|v}[T]^{-1}$$

where

$$[T] = [V]'[V]$$

$$[P] = [V]'[U]$$

and

$$S_{uu|v} = ([U]'[U] - [b]'[P])/(n-m)$$

= estimator of the conditional variance of u given $[v]$

Note that

$$S_{yy|x} = a^2 S_{uu|v}$$

= estimator of the conditional variance of y given $[x]$

For the special case of the regression function of biomass y on tree diameter d , the conditional variance of y given d can be assumed to be approximately proportional to d^4 , that is, $a = d^2$. When the regression of biomass y on tree diameter d and height h is desired, the conditional variance of y given d can be assumed to be approximately proportional to $d^4 h^2$, that is, $a = d^2 h$.

The method of least or weighted least squares in its standard form above assumes implicitly that the variables y and $[x]$ are measured in the interval or ratio scale. When some of these variables are qualitative and their measurement scale is nominal or ordinal, one may still use the least or weighted least squares; but then, these variables must be transformed first to vectors of dummy variables. These procedures are not specifically needed here, even though they can be applied with the methodology described in this paper. The interested people are referred to Cunia (1986c), among others.

To obtain valid estimators $[b]$ and $[S_{bb}]$, certain basic assumptions of the least or weighted least squares must be satisfied first, at least approximately. For a more detailed discussion of the relationship between estimators and basic assumptions of the least squares regression method, the reader is referred to papers by Cunia (1979a, b, 1986b). This is not repeated here. Of interest in the present study is only the

assumption that the sample values y are statistically uncorrelated. This assumption is automatically satisfied when the trees are selected by simple random sampling. However, this is seldom, if ever the case; it is much more efficient to select the sample trees in clusters or by methods other than simple random sampling. As the computer packages used in the least squares estimation of the biomass regression functions assume that the tree biomass values are uncorrelated, valid questions may be raised about the validity of the estimators $[b]$ and $[S_{bb}]$. To insure their validity, appropriate modifications must be made to the standard least squares method; modifications that take into account the effect of the method by which the sample trees were selected.

The objectives of the present paper are those of presenting two modifications of the least or weighted least squares; one that can be applied when the sample trees are selected by cluster sampling, the other when the data of the sample trees are selected by a two-phase or double sampling design. The basic discussions on these two modifications can be found in Cunia (1981, 1982), Briggs and Cunia (1982) and Cunia and Michelakackis (1983). Using simulation techniques, these modified least squares procedures were later tested by Gillespie and Cunia (1986) and Michelakackis and Cunia (1986), among others.

Sample Tree Selection by Cluster Sampling Method

The method of cluster sampling as considered here can be simply described as the selection, by some random procedure of groups or clusters of sample trees (rather than individual trees) from a given forest area of interest. Usually, these clusters consist of trees growing within sample plots of fixed area or trees counted by relascope from fixed points in the forest. They may also consist of subsamples (fixed number or percentages) of trees selected from randomly selected plots or points.

Ideally, this requires (i) a prior subdivision of the tree population into non-overlapping plots of fixed area, and if some conditions are satisfied, overlapping plots or relascope points, (ii) a random sample of plots selected by some statistical sampling procedure and possibly (iii) a random subsample of trees from the plots so selected. Strictly speaking, the cluster sampling (or one-stage cluster sampling) is defined when all the trees of the sample plots are included in the sample and two-stage sampling (or two-stage cluster sampling) when each sample plot is subsampled. For convenience, both methods will be known here simply as cluster sampling.

The way the cluster sampling is actually applied to selection of sample trees for construction of biomass tables is, however, somewhat different. Points are selected in the forest, more or less at random and an arbitrary number of sample trees are selected, by some random, arbitrary or subjective procedure from the forest area around these points. These trees are measured for the characteristics of interest (diameter, height, species, site quality, etc. and

biomass of various tree components) and they constitute the sample trees from which the biomass regression function is being estimated.

The main advantage of cluster sampling is the large decrease in the average sampling costs per tree; as more and more trees are measured from the same location, there is less and less ground movement of the logging equipment and crew. It is also more convenient to work in a few areas than go from tree to tree all over the forest. But there are also disadvantages. One major disadvantage is that the average amount of information per tree decreases with the increase of the average number of sample trees per cluster; trees growing close to each other differ less among themselves than trees growing farther apart. A second major disadvantage is that, unless the least squares techniques are appropriately modified, the validity of the statistical inferences made under the standard assumptions of the least squares method are highly questionable.

Several such modifications have been proposed and applied to samples of trees. A first modification makes use of the theory of ratio estimators as described in standard textbooks on sampling techniques. It is described, and the results of its application discussed by Kotimaki and Cunia (1981), Cunia and Gillespie (1985) and Gillespie and Cunia (1986). This modification is not considered here. A second modification makes use of linear regression techniques with dummy variables. It is described by Cunia (1986b) but, to my knowledge, it has never been applied. It is also not considered here. A third, probably the best modification is to make use of the generalized least squares method. It is a complex procedure, to my knowledge has never been applied to sample trees selected by cluster sampling and is of no further interest here. The last modification, and this is the modification we shall discuss in this paper, uses the ordinary weighted least squares applied to a set of appropriately defined cluster variables; the usual techniques are applied to individual tree variables.

Modification of the Least Squares Regression Method: Trees Selected by Cluster Sampling

Let us make the following, more formal assumptions.

(1) The sample trees are selected in clusters (plots) and the trees are measured for the dependent variable y (biomass of some tree component) and the independent variables $x_1, x_2, \dots, x_m$ (tree measurements other than biomass).

(2) The true regression function of y on $[x]$ is of the form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

and the conditional variance of y given $[x]$ is proportional to some known tree value a^2 , that is

$$\sigma_{yy|x} = a^2 \sigma^2$$

where a^2 but not σ^2 is known.

(3) The sample values y within a cluster may be correlated, but the values y belonging to different clusters are uncorrelated.

There are also additional implicit assumptions not mentioned here as, for example, that conditional probabilities distribution of y given $[x]$ is normal, that the variables x are fixed variables measured without error, that the number of clusters is greater than the number m of variables, that we deal with static populations that do not change with time, etc.

Let us now sum up, for each cluster, the tree variables $y, x_1, x_2, \dots, x_m$, and write the new cluster variables

$$t = \Sigma y, s_1 = \Sigma x_1, s_2 = \Sigma x_2, \dots, s_m = \Sigma x_m$$

where Σ means summation over the trees of a given cluster. For example, if the cluster is a plot and y = tree biomass, $x_1 = 1$, x_2 = tree diameter d and $x_3 = d^2$, the new cluster variables are

$$\begin{aligned} t &= \Sigma y = \text{plot biomass} \\ s_1 &= \Sigma x_1 = \text{number of trees per plot} \\ s_2 &= \Sigma x_2 = \text{sum of tree diameters } d \text{ per plot} \\ s_3 &= \Sigma x_3 = \text{sum of squared diameters } d^2 \text{ per plot} \end{aligned}$$

As the regression values y of the biomass of the trees of a given plot can be written as

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

we can also write the regression value of the plot biomass as

$$\begin{aligned} \hat{t} &= \Sigma \hat{y} = \Sigma (\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) \\ &= \beta_1 s_1 + \beta_2 s_2 + \dots + \beta_m s_m = [\beta]'[s] \end{aligned}$$

Although this looks very much like the true regression of t on $[s]$, it is not. The true regression of the plot biomass t on plot variables $[s]$ is defined as the conditional expected value of t given $[s]$. Its form is not necessarily the same as that of the true regression of tree biomass y on tree variables $[x]$. Various subdivisions of the population of trees into sets of plots yields various regressions of t . The fact that the regression of y on $[x]$ is linear does not mean that the regression of t on $[s]$ is also linear.

Let us consider now the total biomass of the given forest population. Obviously, the total plot biomass is equal to the total tree biomass and also equal to the total regression values of the tree biomass. This means that we can write

$$\begin{aligned} \Sigma y &= \Sigma \hat{y} = \Sigma \hat{t} = \Sigma t \\ &= \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2 + \dots + \beta_m \Sigma x_m \\ &= \beta_1 \Sigma s_1 + \beta_2 \Sigma s_2 + \dots + \beta_m \Sigma s_m \end{aligned}$$

and we can view the function $t = [\beta]'[s]$ as a model for the regression function of plot biomass on plot variables $[s]$. This in turn implies that we can use the sample plot data to estimate $[\beta]$. Furthermore, because t is the sum of several variables y with variances $a^2 \sigma^2$, it is reasonable to assume that the conditional variance of t is at least proportional, if not equal to the sum $\sigma^2 \Sigma a^2$ of variances, where again Σ means summation over the trees of a given plot.

Let us reverse somehow the process and make the following assumptions.

(1) The true regression function of the plot biomass t on plot variables $s_1, s_2, \dots, s_m$ is of the linear form

$$\hat{t} = \beta_1 s_1 + \beta_2 s_2 + \dots + \beta_m s_m = [\beta]'[s]$$

and the conditional variance of t given $[s]$ is proportional to the known sum of tree values Σa^2 .

(2) The sample plot values t are uncorrelated random variables.

(3) The probability distribution of t given $[s]$ is normal; this assumption is needed only for testing null hypotheses or calculating interval estimates.

Then, the weighted least squares estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ are the least squares estimates of the regression of the transformed variables $u = t / \sqrt{\Sigma a^2}$ on $v_1 = s_1 / \sqrt{\Sigma a^2}, v_2 = s_2 / \sqrt{\Sigma a^2}, \dots, v_m = s_m / \sqrt{\Sigma a^2}$. That is, if we define the matrices of sample plot values u and $v_1, v_2, \dots, v_m$ as $[U]$ and $[V]$ and the matrices of the sums of crossproducts as $[T] = [V]'[U]$ and $[P] = [V]'[U]$, then

$$[b] = [T]^{-1}[P]$$

$$[S_{bb}] = S_{uu}|_v [T]^{-1}$$

where

$$S_{uu}|_v = ([U]'[U] - [b]'[P]) / (q - m)$$

where q is the number of sample clusters.

Note what we are doing. We assume a linear regression function $t = [\beta]'[s]$ of the plot biomass t on plot variables $s_1, s_2, \dots, s_m$ and, using the sample plot data, we calculate the estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$. The total biomass in the forest area can now be written successively as the following expressions

$$\Sigma t = \Sigma \hat{t} = \Sigma [\beta]'[s]$$

This is due to a property of the linear least squares regressions; the sum of regression values is equal to the sum of actual values.

$$\begin{aligned} \Sigma \hat{t} &= \beta_1 \Sigma s_1 + \beta_2 \Sigma s_2 + \dots + \beta_m \Sigma s_m \\ &= \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2 + \dots + \beta_m \Sigma x_m, \text{ because} \\ &\quad \text{the variables } s_i \text{ are defined as the sums} \\ &\quad \text{of the variables } x_i \text{ in a given plot} \\ &= \Sigma (\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) = \Sigma [\beta]'[x] \end{aligned}$$

$= \Sigma \hat{y} = \Sigma y$, because sum of the tree biomass (Σy) is identically equal to the sum of the plot biomass ($\Sigma \hat{y}$).

The vector $[\beta]$ of the function $\hat{y} = [\beta]'[x]$ is not the vector of coefficients of the regression function of tree biomass on $[x]$; it is the vector of the regression function of plot biomass on $[s]$. Nevertheless, the function $\hat{y} = [\beta]'[x]$ can serve as an approximation or a substitute for the true regression function of the tree biomass, and $\hat{y} = [b]'[x]$ may serve as an estimate of this regression function. Whether $\hat{y} = [b]'[x]$ is a good estimate of the true regression function of y on $[x]$ depends, among other things on how close the vector $[\beta]$ of the plot regression is to the vector $[\beta]$ of the tree regression. Only experience with actual forest biomass data will show whether the proposed modified procedure above yields acceptably good results.

It may be of interest to show that this modified least squares method makes sense when applied within the specific forest inventory design having plots, not trees as the sampling units. Consider, for example, the two-phase or double sampling design consisting of (i) a first phase, relatively large, simple random sample of plots whose trees are measured for the variables $x_1, x_2, \dots, x_m$ and (ii) a second phase, relatively small, simple random sample of plots whose trees are measured for the biomass y in addition to the variables $x_1, x_2, \dots, x_m$ of the first phase. The plots of the first phase provide plot measurements $s_1, s_2, \dots, s_m$ and estimates $[\bar{s}]' = [\bar{s}_1 \ \bar{s}_2 \ \dots \ \bar{s}_m]$ of the corresponding population means and $[S_{\bar{s}\bar{s}}]$ of the covariance matrix of $[\bar{s}]$. The plots of the second phase provide estimators $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ associated with the biomass regression function

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

by the modified procedure presented above.

Then, the estimate of the mean biomass per plot is

$$\bar{t} = [b]'[\bar{s}]$$

and an estimate of the variance of $\bar{t}$ is

$$S_{\bar{t}\bar{t}} = [b]'[S_{\bar{s}\bar{s}}][b] + [\bar{s}]'[S_{bb}][\bar{s}]$$

The estimator $\bar{t}$ is known as the double sampling with (multiple linear) regression estimator. It can be shown, see Cunia (1986a) that the same estimator $\bar{t}$ can be obtained by the usual procedure where (i) the biomass of each sample tree of the first phase is calculated by the regression $\hat{y} = [b]'[x]$, (ii) the biomass of each plot of the first phase is calculated as the sum of the biomass $\hat{y}$ of all its trees (this is the same as the value given by the formula $\hat{t} = [b]'[s]$ applied to the variables $s_1, s_2, \dots, s_m$ of the given plot) and (iii) the average $\bar{t}$ is calculated as the average of the plot biomass values $\hat{t}$.

An Illustrative Example

To show how this procedure can be applied we shall consider the following numerical example.

Example 1. Consider the data from the 353 sample trees used by Cunia (1986b) to estimate the following biomass regression,

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

$$= b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x]$$

where

y = total above ground tree biomass (pounds of green weight)

d = tree diameter at breast height (inches)

$$[b]' = [5.1818118 \quad -25.653078 \quad 12.988357]$$

= estimate of the vector $[\beta]$ of regression coefficients

and

$$[S_{bb}] = \begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

= estimate of the covariance matrix of $[b]$.

These trees have been assumed to have been selected by simple random sampling in order for Cunia (1986b) to illustrate the application of the ordinary weighted least squares method. In reality they were selected by simulated cluster sampling; more specifically by a two-stage simulated sampling procedure. In the first stage, 30 one-fifth acre sample plots were selected at random and without replacement, in the second stage 30 percent of their trees were selected, again by simple random sampling without replacement. As there is a total of 353 trees in the sample, the average number of sample trees per cluster is about 12; the number vary from a low of 1 to a high of 22 trees.

To apply the modified least squares procedure shown here we start with the grouping of the trees in clusters by plot number and the calculations of the following plot variables

t = plot biomass (Σy)

s_1 = number of trees ($\Sigma 1$) per plot

s_2 = sum of tree diameters (Σd) per plot

s_3 = sum of squared diameters (Σd^2) per plot

and

$\Sigma d^4 = a^2$ = the value, such that the conditional variance of t given $[s]$ is assumed proportional to.

Note that the ordinary weighted least squares method used by Cunia (1986b) assumes that the conditional variance of y given d is proportional to d^4 .

We continue with the calculation of the new transformed variables $u = t/\sqrt{\Sigma d^4}$ and $v_i = s_i/\sqrt{\Sigma d^4}$, $i=1,2,3$. The plot variables t, s_1, s_2, s_3 ,

$a = \sqrt{\sum d^4}$, u , v_1 , v_2 and v_3 are listed in Table 1. For convenience, some of the values were reported with a limited number of significant digits, even though, to calculate them the computer used double precision.

The matrices of the sums of cross products are

$$[T] = [V]'[V] = \begin{bmatrix} .04887035 & .38092564 & 3.2785336 \\ .38092564 & 3.0330882 & 26.811322 \\ 3.2785336 & 26.811322 & 245.29889 \end{bmatrix}$$

$$[P] = [V]'[U] = \begin{bmatrix} 33.258523 \\ 273.51112 \\ 2520.8582 \end{bmatrix}$$

and the inverse of $[T]$ is

$$[T]^{-1} = \begin{bmatrix} 12250.429 & -2696.1925 & 130.96286 \\ -2696.1925 & 603.15120 & -29.888939 \\ 130.96286 & -29.888939 & 1.5205770 \end{bmatrix}$$

As $[U]'[U] = 26860.109$, we finally determine the statistics of interest

$$[b]' = [131.33607 \quad -48.59752 \quad 13.833052]$$

$$S_{uu|v} = 33.809815$$

$$[S_{bb}] = \begin{bmatrix} 414184.73 & -91157.769 & 4427.8299 \\ -91157.769 & 20392.431 & -1010.5395 \\ 4427.8299 & -1010.5395 & 51.410428 \end{bmatrix}$$

At first sight these values look quite different from the values $[b]$ and $[S_{bb}]$ calculated by the ordinary weighted least squares method. However, we never work with individual values of the regression coefficients or their error; we always work with linear combinations. For example, if we wish to calculate the regression estimate of the tree biomass of a 10 inch diameter tree (that is, to calculate the estimate of the average biomass of all 10 inch trees in the forest) we use the formula

$$y = b_1 + 10b_2 + 100b_3$$

Using the ordinary weighted least squares as applied by Cunia (1986b) we obtain the value

$$y = 5.18 - (10)(25.653) + (100)(12.9884) = 1047 \text{ pounds}$$

The modified weighted least squares method presented here yields the value

$$y = 131.34 - (10)(48.600) + (100)(13.8331) = 1029$$

The difference is about 2 percent, well within the inherent sampling error.

The regression estimates and their 95 percent confidence limits, as calculated by the ordinary and the modified least squares procedures are shown in Table 2. Note that the 95 percent confidence limits are defined by the formula

$$[b]'[x] \pm 2\sqrt{[x]'[S_{bb}][x]}$$

where

$$[x]' = [1 \quad d \quad d^2]$$

for $d = 4, 5, \dots, 26$

For example, if $d = 10$, $[x]' = [1 \quad 10 \quad 100]$ and the 95 percent confidence limits become

$$1047.487 \pm 39.813$$

for a lower and upper limit of 1007.674 and 1087.300 respectively for the ordinary least squares and

$$1028.666 \pm 188.294$$

for a lower and upper limit of 840.372 and 1216.960 respectively for the modified least squares.

A look at Table 2 shows that the confidence limits calculated by the modified least squares are much wider. We did expect to have them wider but not by that much. One of the reasons may be the sampling error; we happened to work with a highly unusual sample. Another reason may be that the simulated tree population from which the sample trees were selected is not representative of the real life populations; we have, in the simulated population a much larger variance "between" clusters (relative to the variance "within" clusters) than what we ordinarily have in real life. A third reason is the fact that the two sets of confidence limits refer to two different things; for the ordinary least squares the limits refer to the average biomass per tree for the class of all population trees that have the same diameter, while for the modified least squares they refer to the average biomass per plot (cluster) of all population plots that have one tree of the given diameter. As our sample has no such plot (and plots with one tree only are very rare), the biomass table constructed by the modified least squares contains regression estimates calculated for values of the independent variables s that fall outside the sample data.

It may be of interest to see the impact of the least squares estimation procedure on the estimates of the mean biomass per acre and its error when the biomass regression is applied to an actual forest inventory data. Let us refer to Example 1 of Cunia (1986a) where the ordinary weighted least squares biomass regression was applied to a sample of 926 one-fifth acre plots. He has then found the following results

$$w = 120480 \text{ pounds}$$

= estimate of the mean biomass per acre

$$S_{ww}^{(1)} = 12415782 = \text{estimate of the variance of } w \text{ when the error of the biomass regression is ignored}$$

$$S_{ww} = 19059566 = \text{estimate of the variance of } w \text{ when the error of the biomass regression is accounted for}$$

As the reader can verify, 65.1 percent of the variance is associated with the error of the sample plots and 34.9 percent is associated with the error of the biomass regression.

Table 1 - The plot variables t = biomass (pounds), s_1 = number of trees, s_2 = sum of diameters d , s_3 = sum of squared diameters d^2 , $a = \sqrt{2}d^4$, $u = t/a$, $v_1 = s_1/a$, $v_2 = s_2/a$ and $v_3 = s_3/a$ used in Example 1.

Plot	t	s_1	s_2	s_3	a	u	v_1	v_2	v_3
1	11726	13	110.4	1018.98	333.19	35.19	.03902	.3313	3.058
2	795	2	15.0	118.98	92.39	8.60	.02165	.1624	1.288
3	16870	16	151.7	1545.77	437.49	38.56	.03657	.3467	3.533
4	6580	3	33.6	571.58	512.61	12.84	.00585	.0655	1.115
5	12705	12	97.3	854.09	296.55	42.84	.04046	.3281	2.880
6	20753	21	164.2	1605.20	595.35	34.86	.03527	.2758	2.696
7	9859	12	92.2	783.68	278.50	35.40	.04309	.3311	2.814
8	10083	10	86.1	827.35	315.37	31.97	.03171	.2730	2.623
9	355	1	6.6	43.56	43.56	8.15	.02296	.1515	1.000
10	3599	6	47.6	414.86	201.19	17.89	.02982	.2366	2.062
11	1813	2	15.4	133.16	111.01	16.33	.01802	.1387	1.200
12	7902	16	108.9	767.25	205.49	38.45	.07786	.5299	3.734
13	6220	6	57.0	671.34	398.56	15.61	.01505	.1430	1.684
14	16828	22	189.4	1752.98	434.61	38.72	.05062	.4358	4.033
15	16512	18	164.9	1627.29	443.43	37.24	.04059	.3719	3.670
16	20102	22	202.4	2094.80	550.56	36.51	.03996	.3676	3.805
17	15420	11	119.3	1604.55	712.55	21.64	.01544	.1674	2.252
18	4727	14	95.6	695.42	215.76	21.91	.06489	.4431	3.223
19	7270	15	108.9	825.61	229.83	31.63	.06527	.4738	3.592
20	18169	11	134.2	1912.90	767.99	23.66	.01432	.1747	2.491
21	10737	16	123.4	1040.30	321.68	33.38	.04974	.3836	3.234
22	9043	19	163.1	1505.33	400.50	22.58	.04744	.4072	3.759
23	12248	13	109.5	1096.81	404.10	30.31	.03217	.2710	2.714
24	8074	15	101.1	695.29	186.88	43.20	.08027	.5410	3.721
25	327	2	12.0	72.00	50.91	6.42	.03928	.2357	1.414
26	16994	13	135.8	1503.48	465.06	36.54	.02795	.2920	3.233
27	14575	9	99.1	1208.87	469.34	31.05	.01918	.2111	2.576
28	13405	6	75.9	1097.79	592.85	22.61	.01012	.1280	1.852
29	18541	17	160.0	1667.92	482.82	38.40	.03521	.3314	3.455
30	9699	10	96.0	995.70	358.77	27.03	.02787	.2676	2.775

Table 2 - The biomass tables and their 95 percent confidence limits as constructed in Example 1 by the ordinary weighted least squares (OWLS) and modified weighted least squares (MWLS); the above-ground tree biomass is given in pounds of green weight.

diameter inches	OWLS			MLS		
	lower limit	biomass estimate	upper limit	lower limit	biomass estimate	upper limit
4	67	110	154	0	158	541
5	118	202	225	3	234	465
6	304	319	334	223	337	453
7	443	462	482	396	469	542
8	605	631	657	513	628	743
9	794	826	859	655	814	974
10	1008	1047	1087	840	1029	1217
11	1244	1295	1345	1067	1271	1474
12	1500	1568	1635	1323	1540	1757
13	1775	1867	1958	1592	1837	2083
14	2070	2192	2313	1855	2162	2469
15	2385	2543	2701	2106	2515	2924
16	2719	2920	3121	2345	2895	3445
17	3073	3323	3573	2577	3303	4029
18	3446	3752	4057	2803	3738	4674
19	3840	4207	4573	3025	4202	5378
20	4254	4687	5121	3245	4693	6141
21	4687	5194	5701	3462	5211	6951
22	5141	5727	6314	3677	5757	7838
23	5614	6286	6958	3890	6331	8772
24	6108	6871	7634	4102	6933	9763
25	6622	7482	8341	4313	7562	10811
26	7155	8118	9081	4522	8219	11916

Using the biomass regression estimated by the modified weighted least squares method and the method and statistics of Cunia (1986a) example, we find

$$w = 120672 \text{ pounds}$$

$$= \text{estimate of the mean biomass per acre}$$

$$S_{ww}^{(1)} = 12280114 = \text{estimate of the variance of } w \text{ when the error of the biomass regression is ignored}$$

$$S_{ww} = 49544699 = \text{estimate of the variance of } w \text{ when the error of the biomass regression is accounted for}$$

In this case, 24.8 percent of the error (expressed as variance) is due to sample plots and 75.2 percent is due to the biomass regression.

In terms of the 95 percent confidence limits, an expression that is more meaningful to the layman, we have the following

$$(1) \text{ When the ordinary weighted least squares method is applied}$$

$$w \pm 2\sqrt{S_{ww}^{(1)}} = (120480 \pm 7047) \text{ pounds}$$

and

$$w \pm 2\sqrt{S_{ww}} = (120480 \pm 8731) \text{ pounds}$$

$$(2) \text{ When the modified weighted least squares method is applied}$$

$$w \pm 2\sqrt{S_{ww}^{(1)}} = (120672 \pm 7009) \text{ pounds}$$

and

$$w \pm 2\sqrt{S_{ww}} = (120672 \pm 14078) \text{ pounds}$$

As the reader can verify, the effect of the estimation procedure has a negligible effect on the value of the mean biomass per acre estimate, but may have a critical effect on the estimation of its error. Of course, the values above refer to a simulated population and the corresponding values for natural populations may be different. However, these results may show that the estimation may have a critical effect on the values of the estimates of the error.

Sample Tree Selection by Double Sampling

A common method used in the selection of sample trees (for biomass tables construction) consists of two main phases. In the first phase, a relatively small sample of trees is selected by some random procedure, and the trees are measured for diameter d , height h and biomass y . In the second phase, a relatively large sample of trees is selected, again by some random procedure and the trees are measured for diameter d and height h alone; these trees are not measured for biomass y . The data of the trees from the first phase sample are used to estimate a relationship between biomass, diameter and height; the data of the second phase sample trees are used to estimate a relationship between tree height and diameter; and finally, the two relationships are combined, in some way, to estimate the regression function of tree biomass on diameter.

There are two major computational procedures to combine the data from the two phases. They

are described, among others, by Clutter et al (1983) for forest mensurationists and by Cunia (1982) for forest biometricians. Here is a short description of these procedures.

Procedure 1 - The regression function of the tree biomass y on diameter d and height h , say $\hat{y}=r_1(d,h)$, is first estimated from the sample of tree data of the first phase. The regression of tree height h on tree diameter d , say $\hat{h}=r_2(d)$, is estimated next from the data of the second phase sample trees. Then, the estimate of the regression function of the biomass y on diameter alone $\hat{y}$ is defined as

$$\hat{y} = r(d) = r_1(d, \hat{h}) = r_1(d, r_2(d))$$

This means that the average biomass per tree of given diameter d_0 is estimated by the regression value $\hat{y}_0=r_1(d_0, \hat{h}_0)$, where $\hat{h}_0$ is calculated as $\hat{h}_0=r_2(d_0)$, the estimate of the average height of the tree of the given diameter d_0 .

This procedure has been used in forest inventory for a long time. Data from a sample of trees, usually selected from logging operations, are used to construct a standard two-way volume table (by diameter and height) by graphical first and by least squares techniques later. The trees from a second sample, usually selected from the sample plots of the current inventory are measured for diameter and height. Their data are used to estimate a relationship between tree diameter and height, again by graphical or least squares methods. The one-way local volume table (on diameter alone) is finally constructed from the standard two-way volume table as follows: the average volume of a tree of diameter d_0 is estimated as the volume given by the two-way table for (i) the given diameter d_0 and (ii) the height h_0 obtained from the relationship diameter-height of the second sample above.

Procedure 2 - As in Procedure 1 above, the sample trees of the first phase are measured for biomass y , diameter d and height h and their data are used to estimate the regression function $\hat{y}=r_1(d,h)$. The second phase sample trees are measured for diameter d and height h . Instead of estimating a relationship diameter-height, however, the one way tree biomass table is constructed as follows. First, the biomass y of each individual tree of the second phase sample is estimated by the regression function $\hat{y}=r_1(d,h)$ of the first phase. And then, the regression function $\hat{y}=r(d)$ is estimated from the measured values d and the estimated values $\hat{y}$ of the second phase trees by the usual (graphical or) least squares techniques.

While the procedures for the estimation of the volume or biomass regressions above were known for a long time (and they are also of common use today) little was done to estimate the error of these regressions. An attempt was made lately by Cunia (1982) who proposed two approaches to estimate the error of the regression function of tree biomass on diameter, one for each procedure above, when (i) the samples of the two phases are statistically independent, (ii) all regression functions are assumed to be linear

and (iii) valid estimates of the covariance matrices of the regression coefficients of $\hat{y}=r_1(d,h)$ and $\hat{h}=r_2(d)$ for Procedure 1 and $\hat{y}=r_1(d,h)$ for Procedure 2 are given.

The approach for Procedure 1 was further described, and its applicability demonstrated, by Cunia and Michelakackis (1983). Using simulation techniques, Cunia and Michelakackis (1986) have shown that this approach leads to valid and reliable estimates of the error. It is this approach we shall discuss here. The other approach devised for Procedure 2 has never been applied, to my knowledge, and will not be considered here.

Modification of the Least Squares Regression Method: Trees Selected by Double Sampling

To simplify the description of the procedure we shall assume that the three regressions are of the form

$$\hat{y} = r_1(d,h) = \alpha_1 + \alpha_2 d^2 h + \alpha_3 d + \alpha_4 h + \alpha_5 d h + \alpha_6 d^2$$

$$\begin{aligned} \text{and } \hat{h} = r_2(d) &= \gamma_1 + \gamma_2 d + \gamma_3 d^2 \\ \hat{y} = r(d) &= \alpha_1 + \alpha_2 d^2 (\gamma_1 + \gamma_2 d + \gamma_3 d^2) \\ &+ \alpha_3 d + \alpha_4 (\gamma_1 + \gamma_2 d + \gamma_3 d^2) \\ &+ \alpha_5 d (\gamma_1 + \gamma_2 d + \gamma_3 d^2) + \alpha_6 d^2 \end{aligned}$$

$$= \beta_1 + \beta_2 d + \beta_3 d^2 + \beta_4 d^3 + \beta_5 d^4 = [\beta]'[x]$$

$$\text{where } \beta_1 = (\alpha_1 + \alpha_4 \gamma_1), \quad \beta_2 = (\alpha_3 + \alpha_4 \gamma_2 + \alpha_5 \gamma_1)$$

$$\beta_3 = (\alpha_2 \gamma_1 + \alpha_4 \gamma_3 + \alpha_5 \gamma_2 + \alpha_6)$$

$$\beta_4 = (\alpha_2 \gamma_2 + \alpha_5 \gamma_3), \quad \beta_5 = \alpha_2 \gamma_3$$

and the definition of $[x]$ is straightforward. We have used, for convenience, different notation to denote the coefficients of the different regressions.

Let us assume that we are given estimates $[a]$ of $[\alpha]$ and $[c]$ of $[\gamma]$ and estimates $[S_{aa}]$ and $[S_{cc}]$ of their covariance matrices. Then, the estimate $[b]$ of $[\beta]$ is easily derived as

$$[b] = \begin{bmatrix} a_1 + a_4 c_1 \\ a_3 + a_4 c_2 + a_5 c_1 \\ a_2 c_1 + a_4 c_3 + a_5 c_2 + a_6 \\ a_2 c_2 + a_5 c_3 \\ a_2 c_3 \end{bmatrix}$$

The objective now is to calculate $[S_{bb}]$, an estimate of the covariance matrix of $[b]$.

It can be shown, see Cunia (1982), Cunia and Michelakackis (1983) and Michelakackis and Cunia (1986) that $[S_{bb}]$ can be defined as

$$[S_{bb}] = [C][S_{aa}][C]' + [A][S_{cc}^*][A]'$$

where

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

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

$[S_{aa}]$ = estimator of the covariance matrix of $[a]$

$[S_{cc}]$ = estimator of the covariance matrix of $[c]$

and

$$[S_{cc}^*] = \begin{bmatrix} 0 & [0] \\ [0]' & [S_{cc}] \end{bmatrix}$$

= estimator of the covariance matrix of the extended vector $[c^*]' = [1 \ c_1 \ c_2 \ c_3]$

It may be interesting to note that, in matrix notation

$$[b]' = [C][a] = [A][c^*]$$

Some of the coefficients a and c may be made equal to zero (when they are not significantly different than zero). Then, the formulae above still apply since the corresponding rows and columns of $[S_{aa}]$ and $[S_{cc}]$ will also be made equal to zero. As an example we shall assume that $a_5 = a_6 = 0$, the case used in the example of the next section. Then

$$[a]' = [a_1 \ a_2 \ a_3 \ a_4 \ 0 \ 0]$$

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

$$[S_{aa}] = \begin{bmatrix} S_{a_1 a_1} & S_{a_1 a_2} & S_{a_1 a_3} & S_{a_1 a_4} & 0 & 0 \\ S_{a_1 a_2} & S_{a_2 a_2} & S_{a_2 a_3} & S_{a_2 a_4} & 0 & 0 \\ S_{a_1 a_3} & S_{a_2 a_3} & S_{a_3 a_3} & S_{a_3 a_4} & 0 & 0 \\ S_{a_1 a_4} & S_{a_2 a_4} & S_{a_3 a_4} & S_{a_4 a_4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and $[c^*]$ and $[C]$ and $[S_{cc}^*]$ remain the same.

An Illustrative Example

To show how the procedure of the previous section can be applied we shall use an example already contained in a paper by Cunia and Michelakackis (1983). For more details the interested reader should refer to it.

Example 2 - Between 1967 and 1969, two hundred and eighty Norway Black Spruce trees were selected by cluster sampling from all over Finland by the Finnish Forest Research Institute and measured, among other things, for their diameter at breast height d (cm), total height h (m) and total biomass (kg). For illustration purposes, the 280 trees were divided into two groups; a first group of 56 trees (the first phase sample) were assumed measured for y , d and h and a second group of the remaining 224 trees (the second phase sample) were assumed measured for d and h alone. For convenience we shall also assume that the sample trees were selected by simple random (not cluster) sampling and that the trees of the first are statistically independent of the second group.

We start with the first phase sample of 56 trees. Assuming that (i) the true regression function of y on d and h is

$$\hat{y} = \alpha_1 + \alpha_2 d^2 h + \alpha_3 d + \alpha_4 h + \alpha_5 dh + \alpha_6 d^2$$

(ii) the conditional variance of y given d and h is proportional to $d^4 h^2$ and (iii) the null hypothesis $\alpha_5 = \alpha_6 = 0$ has been accepted, Cunia and Michelakackis (1983) have found the following statistics (where, for convenience, we have reported them with a relatively small number of significant digits)

$$[a]' = [-3.42213 \quad .0167680 \quad 3.72990 \quad -2.07425 \quad 0 \quad 0]$$

= estimate of the vector $[a]'$ of regression coefficients

and

$$[S_{aa}] = \begin{bmatrix} [S_{aa}^*] & [O_1] \\ [O_1]' & [O_2] \end{bmatrix}$$

= estimate of the covariance matrix of $[a]$,

where

$$[S_{aa}^*] = \begin{bmatrix} 12.6354 & .004429 & -.985188 & -.919102 \\ .004429 & .00000239 & -.000492 & -.000249 \\ -.985188 & -.000492 & .215943 & -.060859 \\ -.919102 & -.000249 & -.060859 & .203787 \end{bmatrix}$$

$$[O_1] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad [O_2] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

The data of the 224 trees of the second phase sample was used to estimate the regression function

of the tree height on diameter. It was assumed that the form of the regression function is

$$\hat{h} = \gamma_1 + \gamma_2 d + \gamma_3 d^2$$

and that the conditional variance of h given d is homogeneous. Cunia and Michelakackis (1983) report the following least squares statistics,

$$[c]' = [1.098167 \quad .862612 \quad -.005809]$$

= estimate of the vector $[Y]'$ of regression coefficients

and

$$[S_{cc}] = \begin{bmatrix} .855041 & -.0931381 & .0021821 \\ -.0931381 & .0107251 & -.00026127 \\ .0021821 & -.00026127 & .000006658 \end{bmatrix}$$

= estimate of the covariance matrix of $[c]$.

We can now apply our procedure to estimate the regression function of the tree biomass on diameter alone and its error. We start with the calculation of the vector $[b]$ of coefficients of the regression

$$\begin{aligned} \hat{y} &= r(d) = r_1(d, r_2(d)) \\ &= b_1 + b_2 d + b_3 d^2 + b_4 d^3 + b_5 d^4 \end{aligned}$$

For convenience, the five elements of $[b]$ are listed individually as

$$b_1 = a_1 + a_4 c_1 = -5.700010$$

$$b_2 = a_3 + a_4 c_2 + a_5 c_1 = 1.940622$$

$$b_3 = a_2 c_1 + a_4 c_3 + a_5 c_2 + a_6 = .0304628$$

$$b_4 = a_2 c_2 + a_5 c_3 = .01446426$$

$$b_5 = a_2 c_3 = -.00009740065$$

To estimate the covariance matrix of $[b]$ we write first the vectors and matrices

$$[c^*]' = [1 \quad 1.098167 \quad .862612 \quad -.005809]$$

$$[S_{cc}^*] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & .855041 & -.0931381 & .0021821 \\ 0 & -.0931381 & .0107251 & -.00026127 \\ 0 & .0021821 & -.00026127 & .000006658 \end{bmatrix}$$

$[a]$ and $[S_{aa}]$ as shown above

$$[C] = \begin{bmatrix} 1 & 0 & 0 & 1.098167 & 0 & 0 \\ 0 & 0 & 1 & .862612 & 1.098167 & 0 \\ 0 & 1.098167 & 0 & -.005809 & .862612 & 1 \\ 0 & .862612 & 0 & 0 & -.005809 & 1 \\ 0 & -.005809 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$[A] = \begin{bmatrix} -3.42213 & -2.07425 & 0 & 0 \\ 3.72990 & 0 & -2.07425 & 0 \\ 0 & .0167680 & 0 & -2.07425 \\ 0 & 0 & .0167680 & 0 \\ 0 & 0 & 0 & .0167680 \end{bmatrix}$$

Now the reader can verify that the previous value of [b] is obtained by the matrix multiplications

$$[b] = [C][a] = [A][c^*]$$

and that the elements of $[S_{bb}]$, denoted here, for convenience, as

$$S_{ij} = \text{covariance of } b_i \text{ and } b_j$$

are the following

$$\begin{aligned} S_{11} &= 14.5413 & , & S_{12} = -2.05253 \\ S_{13} &= -.011749 & , & S_{14} = .0068236 \\ S_{15} &= -.000100030 & , & S_{22} = .0308731 \\ S_{23} &= .000671164 & , & S_{24} = -.000983024 \\ S_{28} &= .000013195 & , & S_{33} = .000130202 \\ S_{34} &= -.000013585 & , & S_{35} = .00000035829 \\ S_{44} &= .000004797 & , & S_{45} = -.00000008545 \\ S_{55} &= .000000001953 \end{aligned}$$

One can estimate now the biomass table and its 95 percent confidence limits by the formulae

$$\hat{y} = [b]'[x]$$

$$S_{\hat{y}\hat{y}} = [x]'[S_{bb}][x]$$

$$\hat{y} \pm 2\sqrt{S_{\hat{y}\hat{y}}}$$

where

$$[x]' = [1 \quad d \quad d^2 \quad d^3 \quad d^4]$$

for $d = 4, 5, \dots$

This table is reported by Cunia and Michelakackis (1983) and, thus, it is not repeated here.

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, Vol. II, W.E. Frayer, (Ed.) Colorado State University, Fort Collins, CO, 1979a.
- Cunia, T. On sampling trees for biomass tables construction: some statistical comments. In: 1979 forest resource inventories workshop proceedings, Vol. II, W.E. Frayer, (Ed.), Colorado State University, Fort Collins, CO; 1979b.
- 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. Error of forest inventory estimates: its main components. In: Proceedings of the Workshop on "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; 1986a.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the Workshop on "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; 1986b.
- Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the Workshop on "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; 1986c.
- Cunia, T.; Gillespie, A.J. Cluster sampling and construction of biomass tables: results of a simulation study. In: Proceedings, third annual southern forest biomass workshop, March 12-14, 1985; University of Florida, Gainesville, FL; 1985.
- 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; 1983.
- Gillespie, A.J.; Cunia, T. Estimation of tree biomass tables by cluster sampling: results of a simulated study. In: Proceedings of the Workshop on "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.
- Kotimaki, T.A.; Cunia, T. Effect of cluster sampling in biomass tables construction: ratio estimators models. Canadian Journal of Forest Research 11: 475-486; 1981.

Michelakackis, J.; Cunia, T. Error of biomass regressions: sample trees selected by double sampling. In: Proceedings of the Workshop on "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.

TUTORIAL PAPERS

Error of Sample Plots

ON THE ERROR OF FOREST INVENTORY ESTIMATES:

STRATIFIED SAMPLING AND DOUBLE SAMPLING FOR

STRATIFICATION^{1/}

Tiberius Cunia

Professor of Statistics and Operations Research,
SUNY College of Environmental Science and Forestry,
Syracuse, NY, 13210

When the design of a forest inventory system is stratified sampling and the stratified sampling formulae are used to calculate the error of the estimate of the average biomass per acre, it is common to ignore the error of the biomass regression function used to estimate the sample tree and plot biomass. The approach described in an earlier paper by Cunia to combine the error of biomass regressions with the error from sample plots, when estimates of average biomass per acre are calculated, is extended from simple random to stratified random sampling. As sometimes the stratum sizes are not known but estimated by a second sample we have also considered the double sampling for stratification design.

Introduction

Cunia (1986a) has proposed an approach to combine the error of the biomass regression function with the error of the forest inventory sample plots (or Bitterlich sample points) when inferences are made about the reliability of the biomass estimates per unit area. This approach requires that the estimators be of the form

$$w = b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]'[z]$$

where $[b]$ is the vector of the coefficients of the biomass regression and $[z]$ is a vector of statistics calculated from the data of the sample plots and points. We implicitly assume here that (i) the true regression function of tree biomass on $[x]' = [x_1 \ x_2 \ \dots \ x_m]$ is of the linear form $\hat{y} = [\beta]'[x]$, (ii) the vector $[z]$ is defined so that, the product $[\beta]'[u_z]$, where $[u_z]$ is the expected value of $[z]$, is close, if not identically equal to the parameter of interest μ and (iii) the vectors $[b]$ and $[z]$ are statistically independent. The variance of w is estimated by the approximate formula

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

where $[S_{zz}]$ and $[S_{bb}]$ are the estimates of the covariance matrices of $[z]$ and $[b]$ respectively.

^{1/} Paper based on a set of lecture notes "On the error of biomass estimates in forest inventories: Part 2: The error component from sample plots". Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

In this formula, the first component of S_{ww} may be viewed as an expression of the error due to the sample plots and the second component may be viewed as an expression of the error due to the biomass regression function.

The definition of $[z]$ depends on (i) the sampling design by which the plots or points are selected, (ii) the specific parameter μ one wishes to estimate and (iii) the definition of the independent variables $x_1, x_2, \dots, x_m$ of the regression function. In his paper, Cunia (1986a) makes the assumptions that (i) the sample plots are selected by simple random sampling (or by a systematic sampling method that is equivalent to simple random sampling), (ii) the parameter to estimate is μ = average biomass per acre, and (iii) the definition of the statistics z is based on the plot variables $s_1, s_2, \dots, s_m$ defined as the averages of $x_1, x_2, \dots, x_m$ expressed on a "per acre" basis.

For example, let us assume that the estimate of the regression function of the tree biomass y on the tree diameter d is of the parabolic form

$$\hat{y} = b_1 + b_2 d + b_3 d^2 = b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x]$$

where the definition of x_1, x_2 and x_3 is obvious. Then, for Σ meaning summation over all the trees of a plot, the plot variables are defined as

$$s_1 = (\Sigma x_1)/\text{plot area} \\ = \text{number of trees per acre}$$

$$s_2 = (\Sigma x_2)/\text{plot area} \\ = \text{sum of diameters per acre, and}$$

$$s_3 = (\Sigma x_3)/\text{plot area} \\ = \text{sum of squared diameters per acre.}$$

When, instead of a sample plot we have a Bitterlich sample point, there are similar definitions for the point variables s_1, s_2 and s_3 ; these are not repeated here. For more details, the interested reader is referred to Cunia (1986a).

Let us assume now that the plots (or points) are selected by simple random sampling and the parameter to estimate is μ , the average biomass per acre. Then, the elements of $[z]$ are defined as

$$z_1 = \bar{s}_1 = \Sigma s_1/n = \text{estimate of the average "number of trees" per acre,}$$

$$z_2 = \bar{s}_2 = \Sigma s_2/n = \text{estimate of the average "sum of diameters" per acre, and}$$

$$z_3 = \bar{s}_3 = \Sigma s_3/n = \text{estimate of the average "sum of squared diameters" per acre}$$

where Σ means summation over the n plots (or points) in the sample.

If the sample covariance of s_i and s_j , $i, j = 1, 2, 3$ is defined as usual by the formula

$$s_{s_i s_j} = \Sigma (s_i - \bar{s}_i)(s_j - \bar{s}_j)/(n-1)$$

and the estimate of the covariance matrix of s_1 , s_2 and s_3 is denoted by $[S_{ss}]$, then, the estimate of the covariance matrix of $[z]$ is

$$[S_{zz}] = [S_{ss}]/n$$

Consequently, the estimate of μ and that of its error is

$$w = [b]'[z] \text{ and } S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

where $[S_{bb}]$ is the estimate of the covariance matrix of $[b]$.

It is the objective of this paper to extend the approach described by Cunia (1986a) from simple random to stratified random sampling; that is, to define a stratified sampling estimator of the average biomass per acre μ and to propose a formula to estimate its error. As sometimes the stratum size is not known without error we shall also consider the case where the size of strata are estimated from a second sample. We shall assume that the reader is familiar with the previous paper by Cunia (1986a) and he understands the definition of the plot variables $s_1, s_2, \dots, s_m$. We shall also assume that the reader is familiar with the theory of the methods of stratified random sampling and double sampling for stratification as described, among others, by Cochran (1977).

Stratified Sampling Applied to Forest Inventory

Consider a forest area subdivided into L non-overlapping and exhaustive strata, where A_h = area (in acres) of stratum h , A = area of the entire forest and $Q_h = A_h/A$ = relative size of stratum h . A statistically independent simple random sample of plots of size $n_h > 2$ is selected from each stratum h and, for notational convenience, we shall assume that n_h is small with respect to the size of the stratum h so that, the effect of the finite population correction factor can be ignored. If, for Σ meaning summation over the n_h plots of stratum h , we define

v_{hk} = biomass "per acre" of plot k in stratum h

$\bar{v}_h = (\Sigma v_{hk})/n_h$ = estimator of the average biomass per acre of stratum h , and

$s_{vv}^{hh} = \Sigma (v_{hk} - \bar{v}_h)^2 / (n_h - 1)$ = estimator of the variance of v_{hk} within stratum h

then, for Σ meaning summation over the L strata,

$\bar{v} = \Sigma Q_h \bar{v}_h$ = stratified sample mean
= estimator of the mean biomass per acre for the entire forest area, and

$S_{\bar{v}\bar{v}} = \Sigma Q_h^2 s_{vv}^{hh} / n_h$
= estimator of the variance of $\bar{v}$

When the variance of $\bar{v}$ above is calculated, the error of the biomass regression function (used to calculate v_{hk}) is ignored. To take it

into account we shall write w as the estimator

$$\begin{aligned} w &= \Sigma Q_h \bar{v}_h = \Sigma Q_h (\Sigma v_{hk} / n_h) \\ &= \Sigma Q_h (b_1 \Sigma s_{1hk} + b_2 \Sigma s_{2hk} + \dots + b_m \Sigma s_{mhk}) / n_h \\ &= \Sigma Q_h (b_1 \bar{s}_{1h} + b_2 \bar{s}_{2h} + \dots + b_m \bar{s}_{mh}) \\ &= b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]'[z] \end{aligned}$$

where the first Σ means summation over stratum h , the second Σ means summation over the values s_{hk} of plots k within stratum h , and $z_i = \Sigma Q_h \bar{s}_{ih} =$ stratified mean of the variables s_i , $i=1,2,\dots, m$.

The estimator of the variance of z_i is given by the usual stratified sampling formula

$$S_{z_i z_i} = \Sigma Q_h^2 s_{ii}^{hh} / n_h$$

while the estimator of the covariance of z_i and z_j is a simple extension of the above formula, that is,

$$S_{z_i z_j} = \Sigma Q_h^2 s_{ij}^{hh} / n_h$$

It is more convenient to express these results in a matrix notation as

$$[\bar{s}^h]' = [\bar{s}_{1h} \quad \bar{s}_{2h} \quad \dots \quad \bar{s}_{mh}]$$

$$[S_{ss}^{hh}] = \begin{bmatrix} s_{11}^{hh} & s_{12}^{hh} & \dots & s_{1m}^{hh} \\ s_{21}^{hh} & s_{22}^{hh} & \dots & s_{2m}^{hh} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1}^{hh} & s_{m2}^{hh} & \dots & s_{mm}^{hh} \end{bmatrix}$$

and

$$[S_{zz}] = \Sigma Q_h^2 [S_{ss}^{hh}] / n_h$$

If the estimates of the average biomass per acre for stratum h is required, then

$w_h = [b]'[z^h]$ = estimator of the average biomass per acre μ_h of stratum h

and $S_{w_h w_h} = [b]'[S_{zz}^{hh}][b] + [z^h]'[S_{bb}][z^h]$
= estimator of the variance of w_h

where

$$[z^h] = [\bar{s}^h] \text{ and } [S_{zz}^{hh}] = [S_{ss}^{hh}] / n_h$$

If the estimate of the average biomass per acre for the entire forest area is required then

$w = [b]'[z]$ = estimator of the average biomass per acre

and $S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$
= estimator of the variance of w
when $[z]$ and $[S_{zz}]$ have been defined above.

Example 1 - A forest area of 42336 acres is subdivided into three strata and we shall assume that the size of stratum 1, 2 and 3 is known to be equal, without error, to 17164, 19056 and 6116 acres respectively. Three statistically indepen-

dent samples, one for each stratum, were selected by simple random sampling. There are 82, 112 and 41 sample plots respectively in the samples of stratum 1, 2 and 3. To calculate the biomass we have used the regression function

$$\hat{y} = b_1 + b_2 d + b_3 d^2 = b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x]$$

with the obvious definitions for x_1 , x_2 and x_3 , where y = above ground biomass (pounds of green weight) and d = tree diameter (inches). The original data of the 353 sample trees and the calculations to determine, by weighted least squares, the vector $[b]$ and covariance matrix $[S_{bb}]$ are given in Cunia (1986b). For the convenience of the reader, $[b]$ and $[S_{bb}]$ are shown in Table 1.

The use of this biomass regression function and the fact that the trees are selected in clusters (plots) require the calculation of the three plot variables defined as

$$s_1 = \Sigma x_1 / a = \text{number of trees per acre}$$

$$s_2 = \Sigma x_2 / a = \text{sum of diameters per acre}$$

$$s_3 = \Sigma x_3 / a = \text{sum of squared diameters per acre}$$

The individual values of the $(82 + 112 + 41) = 235$ one tenth acre sample plots are given in Cunia (1986c) and the summary statistics by the stratum needed here, that is, the vectors $[\bar{s}^h]$ of mean values and covariance matrices $[S_{ss}^{hh}]$ for $h = 1, 2, 3$ are given in Table 1.

Using the relationships $[z^h] = [\bar{s}^h]$ and $[S_{zz}^{hh}] = [S_{ss}^{hh}] / n_h$ we can calculate the estimates w_h of the mean biomass per acre and their variances for each stratum h . For convenience, the matrices $[S_{zz}^{hh}]$ are shown in Table 1. The reader can verify that

$$w_1 = [b]'[z^1] = 37427.300 \text{ pounds} = \text{estimate of the mean biomass per acre } \mu_1 \text{ of stratum 1}$$

$$w_2 = [b]'[z^2] = 112268.69 \text{ pounds} = \text{estimate of the mean biomass per acre } \mu_2 \text{ of stratum 2}$$

$$w_3 = [b]'[z^3] = 225726.36 \text{ pounds} = \text{estimate of the mean biomass per acre } \mu_3 \text{ of stratum 3}$$

$$\begin{aligned} S_{w_1} &= [b]'[S_{zz}^{11}][b] + [z^1]'[S_{bb}][z^1] \\ &= 13014675 + 3221354 = 16236029 \\ &= \text{estimate of the variance of } w_1 \end{aligned}$$

$$\begin{aligned} S_{w_2} &= [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2] \\ &= 13124928 + 12147386 = 25272314 \\ &= \text{estimate of the variance of } w_2 \end{aligned}$$

$$\begin{aligned} S_{w_3} &= [b]'[S_{zz}^{33}][b] + [z^3]'[S_{bb}][z^3] \\ &= 53743500 + 49591085 = 103334585 \\ &= \text{estimate of the variance of } w_3 \end{aligned}$$

Table 1 - The basic statistics $[b]$, $[S_{bb}]$, $[\bar{s}^h] = [z^h]$, $[z]$, $[S_{ss}^{hh}]$, $[S_{zz}^{hh}]$ and $[S_{zz}]$ of Example 1.

$[b]'$	$\begin{bmatrix} 5.1818118 & -25.653078 & 12.988357 \end{bmatrix}$
$[S_{bb}]$	$\begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$
$[\bar{s}^1]'$	$\begin{bmatrix} 142.19512 & 743.47195 & 4293.2925 \end{bmatrix}$
$[\bar{s}^2]'$	$\begin{bmatrix} 277.58929 & 1647.1857 & 11786.376 \end{bmatrix}$
$[\bar{s}^3]'$	$\begin{bmatrix} 250.48780 & 1974.9829 & 21179.952 \end{bmatrix}$
$[z]'$	$\begin{bmatrix} 218.78217 & 1328.1538 & 10105.533 \end{bmatrix}$
$[S_{ss}^{11}]$	$\begin{bmatrix} 13753.147 & 69358.124 & 384753.02 \\ 69358.124 & 361014.14 & 2095548.5 \\ 384753.02 & 2095548.5 & 12995719 \end{bmatrix}$
$[S_{ss}^{22}]$	$\begin{bmatrix} 17946.388 & 77549.569 & 234084.91 \\ 77549.569 & 374176.41 & 1614218.3 \\ 234084.91 & 1614218.3 & 13563142 \end{bmatrix}$
$[S_{ss}^{33}]$	$\begin{bmatrix} 15249.756 & 69057.884 & 92042.322 \\ 69057.884 & 337241.41 & 973362.91 \\ 92042.322 & 973362.91 & 15624095 \end{bmatrix}$
$[S_{zz}^{11}]$	$\begin{bmatrix} 167.72130 & 845.83078 & 4692.1100 \\ 845.83078 & 4402.6115 & 25555.470 \\ 4692.1100 & 25555.470 & 158484.38 \end{bmatrix}$
$[S_{zz}^{22}]$	$\begin{bmatrix} 160.23561 & 692.40686 & 2090.0438 \\ 692.40686 & 3340.8608 & 14412.663 \\ 2090.0438 & 14412.663 & 121099.48 \end{bmatrix}$
$[S_{zz}^{33}]$	$\begin{bmatrix} 371.94527 & 1684.3386 & 2244.9347 \\ 1684.3386 & 8225.4001 & 23740.559 \\ 2244.9347 & 23740.559 & 381075.48 \end{bmatrix}$
$[S_{zz}]$	$\begin{bmatrix} 67.794455 & 314.46217 & 1241.5311 \\ 314.46217 & 1572.1751 & 7615.9949 \\ 1241.5311 & 7615.9949 & 58537.684 \end{bmatrix}$

To calculate the estimate of the mean biomass per acre μ for the entire forest area, we must calculate first the relative size of each stratum $h = 1, 2, 3$, that is, to calculate

$$Q_1 = A_1 / A = 17164 / 42336 = .40542328$$

$$Q_2 = A_2 / A = 19056 / 42336 = .45011338$$

$$Q_3 = A_3 / A = 6116 / 42336 = .14446334$$

Then, the numerical values of

$$[z] = Q_1[z^1] + Q_2[z^2] + Q_3[z^3]$$

and

$$[S_{zz}] = (Q_1)^2[S_{zz}^{11}] + (Q_2)^2[S_{zz}^{22}] + (Q_3)^2[S_{zz}^{33}]$$

are calculated and shown in Table 1. Note that $(Q_h)^2$ denotes here the square of Q_h .

Finally,

$w = [b]'[z] = 98316.722$
 = stratified sample estimate of the mean biomass per acre μ of the entire forest area, and

$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$
 = $5919942.4 + 9837039.2 = 15756982$
 = estimate of the variance of w .

The common procedure to calculate the stratified mean and its variance is to (i) apply the tree biomass regression function and calculate the biomass of each individual tree (ii) add the biomass of trees from a given sample plot to obtain the plot biomass, (iii) divide the plot biomass by plot area to obtain the plot biomass per acre v_{hk} , (iv) calculate the stratum average and variance by the formula $\bar{v}_h = \Sigma v_{hk}/n_h$ and $S_{vv}^{hh} = \Sigma (v_{hk} - \bar{v}_h)^2/(n_h-1)$ and finally (v) calculate the stratified mean and its variance by the formulae $\bar{v} = \Sigma Q_h \bar{v}_h$ and $S_{\bar{v}\bar{v}} = \Sigma Q_h^2 S_{vv}^{hh}/n_h$. Applied to our sample data the procedure yields the following statistics

$\Sigma v_{hk} = 3069038.6$	for stratum 1
$= 12574093.3$	for stratum 2
$= 9254780.8$	for stratum 3
$\Sigma (v_{hk})^2 = 2013092800$	for stratum 1
$= 1574846091000$	for stratum 2
$= 2177187330000$	for stratum 3

The individual plot values v_{hk} are listed in Cunia (1986c). Using these sums and sums of squares, one can calculate

$\bar{v}_h = \Sigma v_{hk}/n_h = 37427.300$	for stratum 1
$= 112268.69$	for stratum 2
$= 255726.36$	for stratum 3

and

$S_{\bar{v}\bar{v}}^{hh} = S_{vv}^{hh}/n_h = 13014675$	for stratum 1
$= 13124928$	for stratum 2
$= 53743500$	for stratum 3

As the reader can verify these are the same as the estimates w_1 , w_2 and w_3 found before and the corresponding first variance components, namely $[b]'[S_{zz}^{hh}][b]$ for $h = 1, 2, 3$.

Similarly

$\bar{v} = \Sigma Q_h \bar{v}_h = 98316.722 = w$

and

$S_{\bar{v}\bar{v}} = \Sigma Q_h^2 S_{\bar{v}\bar{v}}^{hh} = 5919942.4$
 = first component of the variance S_{ww} of w .

As expected, the variance of the mean biomass per acre is underestimated, when the common procedure is being applied; the second component associated with the biomass regression is being ignored. However, we did not expect to see this underestimation much more pronounced for the overall stratified mean. The percent of the total variance due to the error of biomass regression function is

$(100)(3221354)/(16236029) = 19.84$ for stratum 1
 $(100)(12147386)/(25272314) = 48.07$ for stratum 2
 $(100)(49591085)/(103334585) = 47.99$ for stratum 3

and

$(100)(9837039)/(15756982) = 62.43$ for the over-all forest area.

This can be explained by the fact that stratification reduces the error due to sample plots but has no real effect on the error due to biomass regression functions.

The Case of Strata with Different Regressions

The approach of the previous section required that the same regression function be used in all strata. It is possible, however, for each stratum h to have its own regression. The regressions may or may not be statistically independent. We shall now present the approach to use when, in general, the regression functions of various strata are different and not statistically independent.

Let us denote by $[b^h]$ and $[S_{bb}^{hh}]$ the vector of regression coefficients and the corresponding covariance matrix of the biomass regression functions of stratum h . Because, for $h \neq k$, $[b^h]$ and $[b^k]$ may not be statistically independent, let us denote their covariance matrix by $[S_{bb}^{hk}]$. In this covariance matrix, the terms associated with the regression coefficient b_i^h are found on row i , while the terms associated with b_j^k are all found in column j . Of course, when $[b^h]$ and $[b^k]$ are independent $[S_{bb}^{hk}] = [0]$, and when $[b^h] = [b^k]$, we have $[S_{bb}^{hh}] = [S_{bb}^{kk}] = [S_{bb}^{hk}]$. Also, we must have $[S_{bb}^{kh}] = [S_{bb}^{hk}]'$.

We shall arrange all vectors $[b^h]$ in the following giant size vector $[B]$ and all covariance matrices $[S_{bb}^{hk}]$ in the following giant size covariance matrix $[S_{BB}]$, both of order mL , or order $(m_1 + m_2 + \dots + m_L)$ when the various regressions have different variables x ,

$$[B] = \begin{bmatrix} [b^1] \\ [b^2] \\ \vdots \\ [b^L] \end{bmatrix} \quad \text{and} \quad [S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [S_{bb}^{12}] & \dots & [S_{bb}^{1L}] \\ [S_{bb}^{21}] & [S_{bb}^{22}] & \dots & [S_{bb}^{2L}] \\ \vdots & \vdots & \ddots & \vdots \\ [S_{bb}^{L1}] & [S_{bb}^{L2}] & \dots & [S_{bb}^{LL}] \end{bmatrix}$$

It is outside the scope of this paper to consider the problem of how to calculate the

covariance matrix $[S_{bb}^{hk}]$. Its value depends on (i) how the sample trees were selected from stratum h and k and (ii) how the data were analyzed and the regressions calculated. For the special case where the sample of trees from the two strata h and k are statistically independent, and some of the coefficients of $[b^h]$ and $[b^k]$ are thought to be common, one can use dummy variables techniques of the type suggested by Cunia (1973, 1986d). Otherwise, one may have to devise some generalized least squares techniques not generally available in textbooks or papers.

The estimates of the mean biomass per acre in stratum h and its variance can be calculated by the formulae of the previous section, namely

$$w_h = [b^h]'[z^h], \text{ and}$$

$$S_{w_h w_h} = [b^h]'[S_{zz}^{hh}][b^h] + [z^h]'[S_{bb}^{hh}][z^h]$$

The estimate of the overall forest area mean biomass per acre μ is calculated by the formula

$$\begin{aligned} w &= Q_1 w_1 + Q_2 w_2 + \dots + Q_L w_L \\ &= Q_1 [b^1]'[z^1] + Q_2 [b^2]'[z^2] + \dots + Q_L [b^L]'[z^L] \\ &= [B]'[Z] \end{aligned}$$

where the giant size vector $[Z]$ is defined as

$$[Z]' = \begin{bmatrix} Q_1 [z^1]' & Q_2 [z^2]' & \dots & Q_L [z^L]' \end{bmatrix}$$

Note that the vectors $[b]$ and $[z]$ above have superscripts, not exponents.

It remains now to show how to calculate the covariance matrix $[S_{ZZ}]$ of $[Z]$, defined here as

$$[S_{ZZ}] = \begin{bmatrix} [S_{ZZ}^{11}] & [S_{ZZ}^{12}] & \dots & [S_{ZZ}^{1L}] \\ [S_{ZZ}^{12}]' & [S_{ZZ}^{22}] & \dots & [S_{ZZ}^{2L}] \\ \vdots & \vdots & \ddots & \vdots \\ [S_{ZZ}^{iL}]' & [S_{ZZ}^{2L}]' & \dots & [S_{ZZ}^{LL}] \end{bmatrix}$$

We have shown how to calculate the covariance matrix $[S_{zz}^{hh}]$ of $[z^h]$. Then, the covariance matrix of $Q_h [z^h]$ is simply the submatrix $[S_{zz}^{hh}] = (Q_h)^2 [S_{zz}^{hh}]$. To evaluate the submatrix $[S_{zz}^{hk}]$ we shall have first a close look at the covariance of $Q_h z_i^h$ and $Q_k z_j^k$. As z_i^h and z_j^k are statistically independent, their covariance is equal to zero and, thus, the covariance of $Q_h z_i^h$ and $Q_k z_j^k$ is also equal to zero. Consequently, for $h \neq k$, we have $[S_{zz}^{hk}] = [0]$, the zero matrix of order m .

The formulae of w and S_{ww} follow now as

$$w = [B]'[Z] = \text{estimator of the mean biomass per acre } \mu \text{ for the entire forest area,}$$

and

$$S_{ww} = [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z]$$

= estimator of the variance of w .

Special case 1 - The same regression function is used in all strata. Then

$$\begin{aligned} [B]' &= \begin{bmatrix} [b]' & [b]' & \dots & [b]' \end{bmatrix} \\ [Z] &= \begin{bmatrix} Q_1 [z^1]' & Q_2 [z^2]' & \dots & Q_L [z^L]' \end{bmatrix} \\ [S_{BB}] &= \begin{bmatrix} [S_{bb}] & [S_{bb}] & \dots & [S_{bb}] \\ [S_{bb}] & [S_{bb}] & \dots & [S_{bb}] \\ \vdots & \vdots & \ddots & \vdots \\ [S_{bb}] & [S_{bb}] & \dots & [S_{bb}] \end{bmatrix} \\ [S_{ZZ}] &= \begin{bmatrix} [S_{zz}^{11}] & [0] & \dots & [0] \\ [0] & [S_{zz}^{22}] & \dots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \dots & [S_{zz}^{LL}] \end{bmatrix} \end{aligned}$$

and it can be shown by complex and cumbersome algebraic manipulations that the formulae

$$w = [B]'[Z] \text{ and}$$

$$S_{ww} = [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z]$$

are identical to those obtained in the previous section. By using the numerical data of the illustrative example of the previous section, the reader can construct the giant size vectors and matrices defined here and verify that the two sets of formulae yield the same results.

Special case 2 - The regression functions of the various strata are all statistically independent, that is, for $h \neq k$, we have

$$[S_{bb}^{hk}] = [0]. \text{ Then}$$

$$w = [B]'[Z]$$

$$= Q_1 [b^1]'[z^1] + Q_2 [b^2]'[z^2] + \dots + Q_L [b^L]'[z^L]$$

$$= Q_1 w_1 + Q_2 w_2 + \dots + Q_L w_L = \Sigma Q_h w_h$$

and

$$S_{ww} = [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z]$$

$$= \Sigma (Q_h)^2 ([b^h]'[S_{zz}^{hh}][b^h] + [z^h]'[S_{bb}^{hh}][z^h])$$

$$= \Sigma (Q_h)^2 S_{w_h w_h}$$

where $S_{w_h w_h}$ is the estimator of the variance of

$w_h = [b^h]'[z^h]$. This formula makes sense, since, for $h \neq k$, the estimators w_h and w_k are statistically independent and by definition of w as $\Sigma Q_h w_h$, the variance formula follows immediately.

Double Sampling for Stratification

The approach outlined in the previous sec-

tions requires that the relative size Q_h be known without error for each stratum $h = 1, 2, \dots, L$. When it is not known, it is sometimes advantageous to select another sample for the sole purpose of estimating Q_h . The procedure is known as two-phase or double sampling for stratification.

More specifically, a first large sample of units is selected from the population of interest by simple random sampling and the units are classified by the stratum they happened to fall into. Let n' be the size of this sample and n'_h be the number of sample units that happened to fall in stratum h . We shall assume that n' is both, sufficiently small with respect to the population size (so that the effect of the finite population correction factor can be ignored) and sufficiently large (so that the probability of obtaining $n'_h = 0$, for some h , is approximately equal to zero). Then, it is known that

$Q_h = n'_h/n' =$ estimator of the relative size of stratum h

$s_{QQ}^{hh} = Q_h(1-Q_h)/n' =$ estimator of the variance of Q_h

and, for $h \neq k$,

$s_{QQ}^{hk} = -Q_h Q_k / n' =$ estimator of the covariance of Q_h and Q_k

Note that, for notational convenience we have used Q_h to denote both, statistic and parameter.

A second, stratified sample is then selected with n_h elements selected from each stratum h . The n_h elements may be a simple random subsample of the n'_h elements of the first sample above, or may be selected completely independent of them. The elements of this second sample are measured for the variables of interest $s_1, s_2, \dots, s_m$ defined in the previous sections.

If we substitute the estimators Q_h for the corresponding true values in the formulae of the stratified sampling of the previous section, we obtain estimators of the mean volume per unit area, say μ and its variance. While the estimator of μ is valid, the estimator of its variance has a major drawback; it assumes that Q_h is known without error and, thus, the error of Q_h is simply ignored. We shall show now how to take this error component into account for both cases (i) when the second phase sample is a subsample of the first and (ii) when the first and second samples are statistically independent.

Case 1: Second Sample Is A Subsample of The First

The double sampling estimator can be written as

$w = [b]'[z] = b_1 z_1 + b_2 z_2 + \dots + b_m z_m$
where

$z_i = \Sigma Q_h \bar{s}_{ih} =$ double sampling estimator of the mean of s_i

The estimator z_i has been discussed by Cochran (1977). Using our notation and assuming that (i) the population size N is so large that the terms

divided by N can be ignored and (ii) the sample size n' is small with respect to the population size N so that we can use the approximation $(N-n')/(N-1) = 1$, we can write Cochran's equation (12.24) of page 333 as

$$s_{z_i z_i} = \Sigma (Q_h)^2 s_{s_i s_i}^{hh} / n_h + \Sigma Q_h (\bar{s}_{ih} - z_i)^2 / n'$$

where Σ is taken over all strata h . This formula can be extended easily and obtain a formula for the covariance of z_i and z_j as

$$s_{z_i z_j} = \Sigma (Q_h)^2 s_{s_i s_j}^{hh} / n_h + \Sigma Q_h (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / n'$$

This is an approximate formula which would suffice in most forest inventory problems. If more exact formulae are needed, the interested reader should refer to Cochran (1977).

As the covariance matrix $[S_{zz}]$ is now defined, we can write the usual variance formula

$$s_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

Case 2: The Two Samples Are Statistically Independent

Let us consider here the more general case where different, not necessarily independent volume regression functions may be used in different strata. As in the previous section, we define the giant size vector

$$[B]' = \begin{bmatrix} [b^1]' & [b^2]' & \dots & [b^L]' \end{bmatrix} \\ \text{and} \\ [Z]' = \begin{bmatrix} [z^1]' & [z^2]' & \dots & [z^L]' \end{bmatrix} \\ = \begin{bmatrix} Q_1 \bar{s}_{11} & Q_1 \bar{s}_{21} & \dots & Q_1 \bar{s}_{m1} & Q_2 \bar{s}_{12} & \dots & Q_L \bar{s}_{mL} \end{bmatrix}$$

The giant size covariance matrix $[S_{BB}]$ of $[B]$ has been defined in the previous section. Because Q_h is no longer known without error, the giant covariance matrix $[S_{zz}]$ would, however, be different. To calculate it, we shall use the following result.

If u_1 and u_2 are two random variables that are statistically independent of two other random variables v_1 and v_2 , then the covariance of $z_1 = u_1 v_1$ and $z_2 = u_2 v_2$ is approximately equal to

$$s_{z_1 z_2} = u_1 u_2 s_{v_1 v_2} + v_1 v_2 s_{u_1 u_2}$$

Applying this rule to the variables $Q_h \bar{s}_{ih}$ and $Q_k \bar{s}_{jk}$ and using the formulae of the variances and covariances of Q_h , Q_k , $\bar{s}_{ih}$ and $\bar{s}_{jk}$ given in this and previous sections, we can state that

(1) the variance of $Q_h \bar{s}_{ih}$, for any i and h , is

$$(Q_h)^2 s_{s_i s_i}^{hh} / n_h + (\bar{s}_{ih})^2 Q_h (1-Q_h) / n'$$

(2) the covariance of $Q_h \bar{s}_{ih}$ and $Q_h \bar{s}_{jh}$ for given h and $i \neq j = 1, 2, \dots, m$ is

$$(Q_h)^2 s_{s_i s_j}^{hh} / n_h + \bar{s}_{ih} \bar{s}_{jh} Q_h (1-Q_h) / n'$$

(3) the covariance of $Q_h \bar{s}_{ih}$ and $Q_k \bar{s}_{jk}$ for $h \neq k = 1, 2, \dots, L$ and $i, j = 1, 2, \dots, m$, is

$$(-\bar{s}_{ih}\bar{s}_{jk} Q_h Q_k / n')$$

since the term $Q_h Q_k \frac{s_{ih}^{hk}}{s_{ij}^{hk}} / n_h = 0$ due to the statistical independence between the plots of different strata.

To calculate $[S_{ZZ}]$ it is, however, much more convenient to use matrix operations. We have already defined the matrix $[S_{SS}^{hh}]$ and we shall now define the product of the mean vectors

$$[\bar{s}^h][\bar{s}^k]' = \begin{bmatrix} \bar{s}_{1h}\bar{s}_{1k} & \bar{s}_{1h}\bar{s}_{2k} & \dots & \bar{s}_{1h}\bar{s}_{mk} \\ \bar{s}_{2h}\bar{s}_{1k} & \bar{s}_{2h}\bar{s}_{2k} & \dots & \bar{s}_{2h}\bar{s}_{mk} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{s}_{mh}\bar{s}_{1k} & \bar{s}_{mh}\bar{s}_{2k} & \dots & \bar{s}_{mh}\bar{s}_{mk} \end{bmatrix}$$

Then, for a given h ,

$$[S_{ZZ}^{hh}] = (Q_h)^2 / n_h [S_{SS}^{hh}] + (Q_h(1-Q_h)/n') [\bar{s}^h][\bar{s}^h]'$$

and for given $h \neq k$,

$$[S_{ZZ}^{hk}] = - (Q_h Q_k / n') [\bar{s}^h][\bar{s}^k]' = [S_{ZZ}^{kh}]'$$

Consequently, the covariance matrix of $[Z]$ can be written as

$$[S_{ZZ}] = \begin{bmatrix} [S_{ZZ}^{11}] & [S_{ZZ}^{12}] & \dots & [S_{ZZ}^{1L}] \\ [S_{ZZ}^{12}]' & [S_{ZZ}^{22}] & \dots & [S_{ZZ}^{2L}] \\ \vdots & \vdots & \ddots & \vdots \\ [S_{ZZ}^{1L}]' & [S_{ZZ}^{2L}]' & \dots & [S_{ZZ}^{LL}] \end{bmatrix}$$

and finally

$$w = [B]'[Z]$$

= double sampling estimator of μ , the mean volume per unit area

and

$$S_{ww} = [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z]$$

= estimator of the variance of w .

Example 2 - Assume that the stratum sizes of 17164, 19056 and 6116 acres of Example 1 were not known without error; they were estimated from a set of photo-points randomly located on aerial photographs of the forest and analyzed as to the stratum they happen to fall. More specifically, let us assume that $n' = 1253$ randomly selected photo-points were classified by stratum and $n'_1 = 508$, $n'_2 = 564$ and $n'_3 = 181$ points were found to fall within stratum 1, 2, and 3 respectively. Then, we calculate first the following statistics.

(1) The relative size of the three strata are estimated as

$$Q_1 = n'_1 / n' = 508 / 1253 = .40542698$$

$$Q_2 = n'_2 / n' = 564 / 1253 = .45011971$$

$$Q_3 = n'_3 / n' = 181 / 1253 = .14445331$$

(2) As the total forest area is known to be equal to $A = 42336$ acres, the area of each stratum A_h is estimated as

$$A_1 = An'_1 / n' = 17164 \text{ acres}$$

$$A_2 = An'_2 / n' = 19056 \text{ acres}$$

$$A_3 = An'_3 / n' = 6116 \text{ acres}$$

These were the areas assumed known without error in Example 1

(3) The variances and covariances of Q_h and Q_k are estimated as

$$S_{QQ}^{11} = Q_1(1-Q_1)/n' = .00019238304$$

$$S_{QQ}^{12} = -Q_1 Q_2 / n' = -.00014564300$$

$$S_{QQ}^{22} = Q_2(1-Q_2)/n' = .00019753548$$

$$S_{QQ}^{13} = -Q_1 Q_3 / n' = -.000046740039$$

$$S_{QQ}^{23} = -Q_2 Q_3 / n' = -.000051892485$$

$$S_{QQ}^{33} = Q_3(1-Q_3)/n' = .000098632524$$

(4) The giant size vector $[B]$ and covariance matrix $[S_{BB}]$, both of order 9 are defined in terms of the subvector $[b]$ and submatrix $[S_{bb}]$ shown in Table 1 as

$$[B] = \begin{bmatrix} [b] \\ [b] \\ [b] \end{bmatrix} \text{ and } [S_{BB}] = \begin{bmatrix} [S_{bb}] & [S_{bb}] & [S_{bb}] \\ [S_{bb}] & [S_{bb}] & [S_{bb}] \\ [S_{bb}] & [S_{bb}] & [S_{bb}] \end{bmatrix}$$

(5) The giant size vector $[Z]$ is defined as

$$[Z] = \begin{bmatrix} [z^1] \\ [z^2] \\ [z^3] \end{bmatrix} = \begin{bmatrix} Q_1[z^1] \\ Q_2[z^2] \\ Q_3[z^3] \end{bmatrix} = \begin{bmatrix} Q_1[\bar{s}^1] \\ Q_2[\bar{s}^2] \\ Q_3[\bar{s}^3] \end{bmatrix}$$

and the giant size covariance matrix $[S_{ZZ}]$ is defined as

$$[S_{ZZ}] = \begin{bmatrix} [S_{ZZ}^{11}] & [S_{ZZ}^{12}] & [S_{ZZ}^{13}] \\ [S_{ZZ}^{12}]' & [S_{ZZ}^{22}] & [S_{ZZ}^{23}] \\ [S_{ZZ}^{13}]' & [S_{ZZ}^{23}]' & [S_{ZZ}^{33}] \end{bmatrix}$$

where the formulae of the submatrices, for given h , are

$$[S_{ZZ}^{hh}] = (Q_h)^2 / n_h [S_{SS}^{hh}] + (Q_h(1-Q_h)/n') [\bar{s}^h][\bar{s}^h]'$$

and for $h \neq k$

$$[S_{ZZ}^{hk}] = - (Q_h Q_k / n') [\bar{s}^h][\bar{s}^k]' = [S_{ZZ}^{kh}]'$$

The numerical values of $\{\bar{s}^h\}$ and $\{s_{ss}^{hh}\}$ are those given in Table 1 and the numerical values of $\{z^h\}$ and $\{s_{zz}^{hk}\}$ are shown in Table 2.

Consequently we can now write

$$w = [B]'[Z] = 98315.308$$

= estimate of the mean biomass per acre

and

$$S_{ww} = [B]'[s_{zz}][B] + [Z]'[s_{BB}][Z]$$

$$= 6525139.8 + 9836818.8 = 16361958.6$$

= estimate of the variance of w.

As the reader can verify, some of these results are close, but not identical to those obtained in Example 1. The two estimates of μ are 98316.722 and 98315.308 and the two variance components associated with the error of the biomass regression function are 9837039.2 and 9836818.8. The differences are due to round-off error; in Example 1 the relative sizes of strata were calculated from the estimated stratum sizes to the nearest acre. On the other hand the estimate of the variance component due to sample plots is somewhat larger in Example 2; 6525139.8 compared to the value 5919942.4 of Example 1. This is due to the fact that in Example 1 the effect of the error of Q_1 , Q_2 and Q_3 was ignored.

Table 2 - The numerical values of the subvectors $\{z^h\}$ and submatrices $\{s_{zz}^{hk}\}$ of Example 2

$$[z^1] = \begin{bmatrix} 57.649738 \\ 301.42358 \\ 1740.6166 \end{bmatrix}, \quad [z^2] = \begin{bmatrix} 124.94841 \\ 741.43076 \\ 5305.2802 \end{bmatrix},$$

$$[z^3] = \begin{bmatrix} 36.183793 \\ 285.29283 \\ 3059.5142 \end{bmatrix}$$

$$[s_{zz}^{11}] = \begin{bmatrix} 31.458403 & 159.36844 & 888.69397 \\ 159.36844 & 830.00163 & 4814.6545 \\ 888.69397 & 4814.6545 & 29596.314 \end{bmatrix}$$

$$[s_{zz}^{12}] = \begin{bmatrix} -5.7487974 & -34.112761 & -244.09259 \\ -30.057779 & -178.35971 & -1276.2463 \\ -173.57325 & -1029.9654 & -7369.8794 \end{bmatrix} = [s_{zz}^{21}]'$$

$$[s_{zz}^{22}] = \begin{bmatrix} 47.686234 & 230.60834 & 1069.7501 \\ 230.60834 & 1212.8417 & 6755.1404 \\ 1069.7501 & 6755.1404 & 51977.059 \end{bmatrix}$$

$$[s_{zz}^{13}] = \begin{bmatrix} -1.6647935 & -13.126143 & -140.76631 \\ -8.7044282 & -68.630476 & -736.00139 \\ -50.265053 & -396.31718 & -4250.1526 \end{bmatrix} = [s_{zz}^{31}]'$$

$$[s_{zz}^{23}] = \begin{bmatrix} -3.6082262 & -28.449230 & -305.09292 \\ -21.410836 & -168.81475 & -1810.3894 \\ -153.20444 & -1207.9476 & -12954.174 \end{bmatrix} = [s_{zz}^{32}]'$$

$$[s_{zz}^{33}] = \begin{bmatrix} 22.982543 & 125.46803 & 693.13551 \\ 125.46803 & 707.59481 & 4330.4116 \\ 693.13551 & 4330.4116 & 35393.175 \end{bmatrix}$$

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

- Cochran, W. G. Sampling Techniques, 3rd Ed. John Wiley and Sons, New York, NY; 1977.
- Cunia, T. Dummy variables and some of their uses in regression analysis. In: Proceedings of the June 1973 meeting of IUFRO Subject Group S4.02, Nancy-France, Vol. 1, T. Cunia, K. Kuusela and A. J. Nash (Eds), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1973.
- Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the workshop on "Tree biomass functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "Tree biomass functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.
- Cunia, T. On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986c.
- Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the workshop on "Tree biomass functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

ON THE ERROR OF FOREST INVENTORY ESTIMATES: TWO-STAGE SAMPLING OF PLOTS⁽¹⁾

Tiberius Cunia

Professor of Statistics and Operations Research
SUNY College of Environmental Science and Forestry
Syracuse, NY, 13210

The error of forest biomass estimates has two main sources, the sample plots where the trees are measured for diameter only and the regression functions that are used to estimate the biomass. Given that the biomass regressions are linear with known error, an approach is suggested to combine this error with the error of the sample plots where the sample plots are selected by a two-stage cluster sampling design.

Introduction

It is common to have forest inventory designs consisting of trees selected in two phases. In the first phase the trees of a sample of plots (or Bitterlich relascope points) are measured for diameter, species and possibly other attributes but not measured for biomass. In the second phase a sample of trees is measured for biomass in addition to diameter, species and possibly other attributes. The trees of the second phase are then used to estimate the regression function of tree biomass on diameter (and possibly other attributes), which applied to trees of the first phase yields estimates of the average biomass per acre. The error of the estimates has two main components, one component associated with each of the two sampling phases above. However, it is common to ignore the second error component (due to the biomass regression function) when the error of the average biomass per acre estimate is calculated.

An approach proposed by Cunia (1965, 1986a) can be used to combine the error from the first phase sample plots with the error from the second phase sample trees. This approach requires that the estimators be of the form

$$w = b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]' [z]$$

where (i) $[b]$ is the second phase sample estimator of the vector of coefficients of the regression of biomass on $x_1, x_2, \dots, x_m$ (functions of various tree attributes other than biomass) assumed to be of the linear form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]' [x]$$

and (ii) $[z]$ is a vector of statistics calculated from the data of the first phase sample plots. The two vectors $[b]$ and $[z]$ are assumed to be statistically independent and the variance of w can be estimated by the approximate formula

$$S_{ww} = [b]' [S_{zz}] [b] + [z]' [S_{bb}] [z]$$

Note that in this formula $[S_{bb}]$ denotes the covariance matrix of $[b]$ and that the right hand side of the equation has two terms; the first term is the variance component associated with the error of the sample plots (of the first phase) and the second term is the variance component associated with the sample trees, or biomass regression (of the second phase).

The definition of the vector $[z]$ depends on (i) the sampling design of the first phase, (ii) the parameter μ one wishes to estimate and (iii) the type of variables x used in the biomass regression function. In his paper, Cunia (1986a) assumes that (i) the sample plots of the first phase are selected by simple random sampling and (ii) the parameter μ is the mean biomass per acre. Given a biomass regression function of the form

$$\hat{y} = b_1 x_1 + b_2 x_2 + \dots + b_m x_m = [b]' [x]$$

the statistics $z_1, z_2, \dots, z_m$ are then defined as the averages of the sample plot variables $s_i = (\sum x_i)/a$, $i = 1, 2, \dots, m$, where a is the plot area (in acres) and $\sum$ is taken over the trees of a given sample plot. Note that s_i is the sum of the variables x_i of the trees from within a given plot, expressed on a per acre basis. For example, if $x_1 = 1$ and $x_2 = d$ (diameter), then

s_1 = number of trees per acre of a given plot,
and s_2 = sum of diameters per acre of a given plot

If, for n = number of sample plots, the sample means, variances and covariances of the plot variables s_i are denoted by

$$\bar{s}_i = (\sum s_i)/n, \text{ and}$$

$$S_{s_i s_j} = \sum (s_i - \bar{s}_i)(s_j - \bar{s}_j)/(n-1)$$

where $i, j = 1, 2, \dots, m$ and $\sum$ is taken over the n sample plots, then

$$z_i = \bar{s}_i \text{ and } S_{z_i z_j} = S_{s_i s_j} / n$$

Assuming now that $[b]$ and $[S_{bb}]$ are given, the estimates w of μ and S_{ww} of the variance of w are given by the formulae above.

The objective of the present paper is to define the vector $[z]$ of statistics when (i) the sample plots are selected by a two-stage cluster sampling design and (ii) the parameter μ of interest is the average biomass per acre. We shall also show how to calculate $[S_{zz}]$ an estimator of the covariance matrix of $[z]$. We shall assume that the biomass regression function is of the linear form and that we are given the vector $[b]$ of re-

⁽¹⁾ Paper based on a set of lecture notes "On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

gression coefficients and the covariance matrix $[S_{bb}]$ of $[b]$. For a description of the methodology to calculate $[b]$ and $[S_{bb}]$ the reader is referred to Cunia (1986b, c) among others.

We shall also assume that the reader is familiar with the two-stage cluster sampling theory as described in standard textbooks as, for example, that by Cochran (1977). If not familiar with the approach to combine the error from sample plots and biomass regressions, the reader is strongly advised to read the Cunia (1986a) paper. In particular, he should understand well the definition of the plot variables $s_1, s_2, \dots, s_m$ for both sample plots of fixed area and relascope sample points. To facilitate the discussion we shall often use sample plots to denote both cases, of plots and points.

Two-stage Cluster Sampling Applied to Forest Inventory

The clusters of trees previously defined as sample plots of fixed area (or Bitterlich sample points) can be grouped into clusters of higher order (clusters of plots) and the sampling can be done in two stages. Let us denote the clusters of plots as primary sampling units and the plots themselves as secondary units. Within this context, the trees constitute the tertiary units. For example, a large forest area (an entire country or one of its political or administrative subdivisions) can be divided first into blocks, divisions or homogeneous stands (the primary units) and each block, division or stand can be further subdivided into a certain number of sample plots (the secondary units). The trees of the original forest population are the tertiary sampling units.

The general two-stage sampling design consists of (i) a first stage where a sample of primary units is selected by some random procedure and (ii) a second stage where a separate subsample of secondary units is selected at random from each of the primary units selected in the first stage. When the units of the original population (the tertiary units) found in all the sample units of the second stage are measured for the variables of interest we have the method of two-stage sampling. But when each secondary unit of the sample of the second stage is subsampled for the tertiary units we have a three-stage sampling design.

Note that strictly speaking, two-stage sampling refers to the selection of the secondary units which are completely measured for the tertiary units found within them. The estimates are calculated for the parameters of the population of secondary units. When the estimates refer to the population of tertiary units one may wish to denote the sampling method as three-stage. It is only a matter of terminology.

To introduce the necessary terminology, notation and formulae, let us assume that (i) the forest area is divided into M homogeneous stands, (the primary units), (ii) each stand h has a known area of A_h acres, $h = 1, 2, \dots, M$ and can be subdivided either into a finite number N_h of non-overlapping

plots of fixed area "a" acres (the secondary units) or into infinitely many overlapping plots of fixed area or Bitterlich sample points of given basal area factor "c", (iii) each sample plot or point k of the stand h contains n_{hk} trees (the tertiary units), $k = 1, 2, \dots, N_h$ (or infinity) and (iv) the diameter of the g -th tree in the k -th plot (or point) of the h -th stand is denoted by d_{hkg} . Then, the two-stage random sampling design considered here can be described as follows.

In the first stage, m stands are selected by simple random sampling without replacement. Each selected stand of known area of A_h acres is subdivided (at least conceptually) into N_h non-overlapping sample plots of "a" acres each, or infinitely many overlapping plots of fixed area or Bitterlich sample points. In the second stage, n_h sample plots or points are selected from each sample stand $h = 1, 2, \dots, m$ by simple random sampling without replacement (or an equivalent systematic sampling design). All n_{hk} trees of the selected hk -th plot or point, $k = 1, 2, \dots, n_h$ are measured for their diameters d_{hkg} , $g = 1, 2, \dots, n_{hk}$.

To introduce the necessary formulae, let us assume that the k -th plot or point of the h -th sample stand is measured for its value y_{hk} = biomass per acre. Then, two-stage cluster sampling theory tells us that, for sufficiently large samples, a slightly biased, but efficient estimator of μ , the average biomass per acre for the entire forest area, is the ratio estimator

$$w = \bar{y}_R = (\sum N_h \bar{y}_h) / (\sum N_h)$$

where $\bar{y}_h$ is the estimator of the average biomass per acre of stand h , say μ_h , calculated by the formula

$$\bar{y}_h = (y_{h1} + y_{h2} + \dots + y_{hn_h}) / n_h$$

and Σ means summation over the sample stands $h = 1, 2, \dots, m$. When N is equal to infinity, the case of the overlapping sample plots of fixed area or Bitterlich sample points, the formula becomes

$$w = \bar{y}_R = (\sum A_h \bar{y}_h) / (\sum A_h)$$

The variance of $\bar{y}_R$ can be estimated by the formula

$$s_{\bar{y}_R \bar{y}_R} = \left(\frac{M-m}{M} \right) \left(\frac{m}{m-1} \right) \left(\frac{\sum N_h^2 (\bar{y}_h - \bar{y}_R)^2}{\sum N_h} \right) / (\sum N_h)^2 + \left(\frac{m}{M} \right) \Sigma \left(\frac{N_h - n_h}{N_h} \right) \left(\frac{\sum_{yy}^{hh} S_{yy}^{hh}}{n_h} \right) / (\sum N_h)^2$$

where Σ means again summation over $h = 1, 2, \dots, m$ sample stands and S_{yy}^{hh} is an estimate of the variance within the sample stand h , that is the sample variance of the n_h plot values y_{hk} , $k = 1, 2, \dots, n_h$ and $h = 1, 2, \dots, m$. When M is large with respect to m , one can make $(M-m)/M = 1$. Also, when m is sufficiently large, one can make $m/(m-1) = 1$. If, for a given h , N_h is large relative to n_h the factor $(N_h - n_h)/N_h$ can be made

equal to 1. Finally, when N_h is infinitely large, as the case may be with overlapping sample plots of fixed area or relascope sample points, one can substitute A_h for N_h . If all four conditions above are satisfied, one can use the following approximate but simpler formula

$$s_{\bar{y}_R \bar{y}_R} = \left(\Sigma A_h^2 (\bar{y}_h - \bar{y}_R)^2 + m \Sigma A_h^2 s_{yy}^{hh} / M n_h \right) / (\Sigma A_h)^2$$

Sometimes the variation within the stands as measured by s_{yy}^{hh} is relatively small compared to the variance between stand means as represented by the squared differences $(\bar{y}_h - \bar{y}_R)^2$. Then, the effect of the second variance component

$$m \Sigma A_h^2 s_{yy}^{hh} / M n_h (\Sigma A_h)^2$$

can be ignored and the variance of $\bar{y}_R$ can be estimated by the formula

$$s_{\bar{y}_R \bar{y}_R} = \Sigma A_h^2 (\bar{y}_h - \bar{y}_R)^2 / (\Sigma A_h)^2$$

All the above formulae of the variance of $\bar{y}_R$ ignore the error of the biomass regression function used for the calculation of the plot values y_{hk} . We shall now show how to include the error of the regression function into the error of the estimate $\bar{y}_R$ of μ . We shall consider first the case of a single biomass regression applied to all trees of all stands and then extend the results to the case of different regressions applied to different stands. To facilitate our discussion, we shall assume that the biomass regression function is of the form

$$\begin{aligned} \hat{y} &= b_1 + b_2 d + b_3 d^2 \\ &= b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x] \end{aligned}$$

where d is the tree diameter and the definition of $[x]$ is obvious.

Case 1 - Same biomass regression function applied to all stands

The biomass per acre of the hk -th plot or point can be defined by the usual formula as

$$y_{hk} = b_1 s_{1hk} + b_2 s_{2hk} + b_3 s_{3hk}$$

where

s_{1hk} = number of trees per acre of the hk -th plot or point

s_{2hk} = average sum of tree diameters per acre of the hk -th plot or point

s_{3hk} = average sum of squared diameters per acre of the hk -th plot or point

If we write

$$\bar{y}_h = b_1 \bar{s}_{1h} + b_2 \bar{s}_{2h} + b_3 \bar{s}_{3h}$$

where $\bar{s}_{1h}$, $\bar{s}_{2h}$, and $\bar{s}_{3h}$ are the averages of the n_{hk} sample values s_{1hk} , s_{2hk} , and s_{3hk} respectively of the sample stand h , we can write the estimate of μ , the average biomass per acre, as

$$w = b_1 z_1 + b_2 z_2 + b_3 z_3 = [b]'[z]$$

where, for Σ meaning summation over the m sample stands h ,

$$z_1 = \Sigma A_h \bar{s}_{1h} / \Sigma A_h = \Sigma N_h \bar{s}_{1h} / \Sigma N_h$$

$$z_2 = \Sigma A_h \bar{s}_{2h} / \Sigma A_h = \Sigma N_h \bar{s}_{2h} / \Sigma N_h$$

$$z_3 = \Sigma A_h \bar{s}_{3h} / \Sigma A_h = \Sigma N_h \bar{s}_{3h} / \Sigma N_h$$

Each z_i value, $i = 1, 2, 3$, is a ratio estimator of the type $\bar{y}_R$ defined above. Its variance formula has been given above. Extending this formula, we can write the covariance of z_i and z_j as

$$s_{z_i z_j} = \left(\frac{M-m}{M} \right) \left(\frac{m}{m-1} \right) \Sigma N_h^2 (\bar{s}_{ih} - z_i)(\bar{s}_{jh} - z_j) / (\Sigma N_h)^2 + \left(\frac{m}{M} \right) \Sigma \left(\frac{N_h - n_h}{N_h} \right) \left(\frac{N_h^2 s_{ij}^{hh}}{n_h} \right) / (\Sigma N_h)^2$$

where Σ is taken over the m sample stands and s_{ij}^{hh} is the sample covariance of the n_h pairs of values s_{ihk} and s_{jkh} of the sample stand h , that is

$$s_{ij}^{hh} = \Sigma (s_{ihk} - \bar{s}_{ih})(s_{jkh} - \bar{s}_{jh}) / (n_h - 1)$$

where now Σ means summation within the stand h over plot $k = 1, 2, \dots, n_h$. When we can approximate the values $(M-m)/M$, $m/(m-1)$ and $(N_h - n_h)/N_h$ by 1, and if N_h is infinitely large for all $h = 1, 2, \dots, m$, the formula above simplifies to

$$s_{z_i z_j} = \left(\Sigma A_h^2 (\bar{s}_{ih} - z_i)(\bar{s}_{jh} - z_j) + m \Sigma A_h^2 s_{ij}^{hh} / M n_h \right) / (\Sigma A_h)^2$$

Finally, when the covariance within clusters is small relative to the covariance between cluster means, one can use the formula

$$s_{z_i z_j} = \Sigma A_h^2 (\bar{s}_{ih} - z_i)(\bar{s}_{jh} - z_j) / (\Sigma A_h)^2$$

Case 2 - Different biomass regression functions applied to different stands

Many times, the various stands are defined in terms of forest type or site for which the trees are of a different shape, and thus, different regression functions have been constructed for different stands. Then, let us assume that

$$[b^h]' = [b_{1h} \ b_{2h} \ b_{3h}]$$

is the regression applied to stand h , that the covariance matrix of $[b^h]$ is as usually denoted as $[s_{bb}^{hh}]$, and that the m vectors of regression coefficients $[b^h]$ have been calculated by a least squares technique (or some of its extensions) such that the covariance matrix of $[b^h]$ with $[b^k]$ denoted here as $[s_{bb}^{hk}]$ can be calculated. If we define the giant size vector $[B]$ as

$$[B]' = [[b^1]' \ [b^2]' \ \dots \ [b^m]']$$

then, the covariance matrix of $[B]$ can be written as

$$[S_{BB}] = \begin{bmatrix} [s_{bb}^{11}] & [s_{bb}^{12}] & \dots & [s_{bb}^{1m}] \\ [s_{bb}^{12}]' & [s_{bb}^{22}] & \dots & [s_{bb}^{2m}] \\ \vdots & \vdots & \ddots & \vdots \\ [s_{bb}^{1m}]' & [s_{bb}^{2m}]' & \dots & [s_{bb}^{mm}] \end{bmatrix}$$

For the special case where the same regression function is applied to stands h and k, then

$$[s_{bb}^{hh}] = [s_{bb}^{kk}] = [s_{bb}^{hk}]$$

while for the case of statistically independent vectors $[b^h]$ and $[b^k]$, we have $[s_{bb}^{hk}] = [0]$.

The biomass per acre of the hk-th plot or point can be defined as

$$y_{hk} = b_{1h} s_{1hk} + b_{2h} s_{2hk} + b_{3h} s_{3hk}$$

where s_{1hk} , s_{2hk} , and s_{3hk} are defined as in Case 1 above. Also, the average biomass per acre of stand h is estimated by

$$\bar{y}_h = b_{1h} \bar{s}_{1h} + b_{2h} \bar{s}_{2h} + b_{3h} \bar{s}_{3h} = [b^h]' [\bar{s}^h]$$

where $\bar{s}_{1h}$, $\bar{s}_{2h}$, and $\bar{s}_{3h}$ are again defined as in Case 1 above. Then, the two-stage sampling estimator of μ can be written as

$$\begin{aligned} w &= (A_1 \bar{y}_1 + A_2 \bar{y}_2 + \dots + A_m \bar{y}_m) / (A_1 + A_2 + \dots + A_m) \\ &= (A_1 / \Sigma A_k) (b_{11} \bar{s}_{11} + b_{21} \bar{s}_{21} + b_{31} \bar{s}_{31}) \\ &+ (A_2 / \Sigma A_k) (b_{12} \bar{s}_{12} + b_{22} \bar{s}_{22} + b_{32} \bar{s}_{32}) \\ &+ \dots \\ &+ (A_m / \Sigma A_k) (b_{1m} \bar{s}_{1m} + b_{2m} \bar{s}_{2m} + b_{3m} \bar{s}_{3m}) \\ &= [B]' [Z] \end{aligned}$$

where

$$\begin{aligned} [Z]' &= [A_1 [\bar{s}_1]'] \quad A_2 [\bar{s}_2]' \quad \dots \quad A_m [s_m]' / \Sigma A_k \\ &= [[z^1]' \quad [z^2]' \quad \dots \quad [z^m]'] \end{aligned}$$

with $\Sigma A_k = A_1 + A_2 + \dots + A_m$, and

$$[z^h]' = (A_h / \Sigma A_k) [\bar{s}_{1h} \quad \bar{s}_{2h} \quad \bar{s}_{3h}]$$

The covariance matrix of $[z^h]$ is calculated as in Case 1 above and denoted as $[s_{zz}^{hh}]$. As the sample plots or points of stand h are selected independently of the sample plots or points of stand k, the covariance matrix of $[z^h]$ with $[z^k]$ is equal to zero; that is $[s_{zz}^{hk}] = [0]$ when $h \neq k$. Consequently

$$S_{ww} = [B]' [S_{zz}] [B] + [Z]' [S_{BB}] [Z]$$

An Illustrative Example

To better see how the formulae of the previous section apply to an actual set of sample data, let us consider the following numerical example.

Example - Assume that a forest area of $A = 42336$ acres is divided into 140 blocks of sizes varying from about 100 to about 500 acres each. We are given that (i) $m = 16$ blocks are selected by simple random sampling without replacement from the $M = 140$ population blocks, (ii) each block so selected is measured for its area, say A_h in acres, (iii) n_h one-tenth acre sample plots are selected by simple random sampling without replacement from the h-th sample block, with the sample size n_h being approximately proportional to A_h , and (iv) the total sample size is $n = 235$. It is implicitly assumed that the blocks and plots are non-overlapping. The values A_h and n_h are shown by block in Table 1.

To calculate the biomass we shall use the regression function already used by Cunia (1986a). This regression is assumed to be of the linear (parabolic) form

$$\begin{aligned} \hat{y} &= \beta_1 + \beta_2 d + \beta_3 d^2 \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = [\beta]' [x] \end{aligned}$$

where y = tree biomass and d = tree diameter, with obvious definitions for x_1, x_2 and x_3 . The estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ are shown in the Cunia (1986a) paper, and, for convenience they are also given below

$$[b]' = [5.1818118 \quad -25.653078 \quad 12.988357]$$

and

$$[S_{bb}] = \begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

This regression function implies that the plot variables are defined as

$$\begin{aligned} s_1 &= \Sigma x_1 / .10 = 10 \Sigma (1) \\ &= \text{number of trees per acre} \end{aligned}$$

$$\begin{aligned} s_2 &= \Sigma x_2 / .10 = 10 \Sigma d \\ &= \text{sum of diameters per acre, and} \end{aligned}$$

$$\begin{aligned} s_3 &= \Sigma x_3 / .10 = 10 \Sigma d^2 \\ &= \text{sum of squared diameters per acre,} \end{aligned}$$

where Σ means summation over the trees of a given plot.

Consider now the sample of plots. The values s_1, s_2 and s_3 are calculated for each plot separately. The individual plot values are listed by Cunia (1986d, Tables 1 and 2) and are not repeated here. For the purpose of this example, the 235 plots are distributed arbitrarily among the 16 blocks, according to the number n_h of plots that they should contain. More specifically, the first $n_1 = 13$ plots were assigned to block 1, the next $n_2 = 15$ plots were assigned to block 2, etc. with the last $n_{16} = 18$ plots assigned to the last block 16.

Using the terminology and notation of the previous section, the secondary units are the

Table 1 - The area A_h (acres), the number n_h of sample plots, the averages of s_1 , s_2 (inches), s_3 (squared inches) and y (pounds) and the variance V of y within blocks by block number h

h	A_h	n_h	$\bar{s}_1$	$\bar{s}_2$	$\bar{s}_3$	$\bar{y}$	V
1	280	13	210.76923	1243.5154	8765.4079	83040.42	4949333039
2	302	15	200.66667	1175.9533	8444.4321	80552.29	3840381639
3	452	22	245.45455	1403.3682	9800.8083	92567.58	5702946485
4	216	10	149.00000	1010.4200	8361.1126	83448.82	6457081439
5	384	18	205.55556	1323.7056	11052.1138	110656.83	6406414888
6	290	14	173.57143	1380.9929	15067.3455	161172.76	4858967459
7	242	12	279.16667	1588.3583	11129.5609	105255.02	3023986664
8	315	15	230.66667	1448.6333	12156.5293	121926.71	7706768567
9	344	17	232.35294	1289.7118	8270.5904	75540.31	2658394967
10	287	14	166.42857	958.6500	6639.4076	62505.07	4666270466
11	472	23	226.52174	1320.5826	9660.5531	92771.50	3502826347
12	130	6	218.33333	1552.1833	13833.9085	140992.82	7053833420
13	212	10	264.00000	1801.9300	16101.5319	164275.39	8978786810
14	274	13	251.53846	1753.6385	15748.0664	160858.71	3564467131
15	319	15	271.33333	1504.5600	9355.6103	84323.41	2743051584
16	382	18	258.88889	1626.2444	12780.5912	125622.22	7090217424

sample plots and the primary units are the blocks. There is a total of $N = 423360$ possible one-tenth acre plots and $M = 140$ blocks. The first stage sample size is $m = 16$ sample blocks and the second stage sample size is $n = 235$ sample plots. The total area of the 16 sample blocks is

$$\Sigma A_h = A_1 + A_2 + \dots + A_{16} = 280 + 302 + \dots + 382 = 4901 \text{ acres}$$

The first step is the calculation of the averages of the variables s_1 = number of trees per acre, s_2 = sum of tree diameters per acre, and s_3 = sum of squared tree diameters per acre for each sample block $h = 1, 2, \dots, 16$ separately. The 16 sets of values $\bar{s}_{1h}$, $\bar{s}_{2h}$, and $\bar{s}_{3h}$ are listed in Table 1.

We continue with the calculation of the vector $[z]$ defined as

$$[z] = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 225.58049 \\ 1390.1177 \\ 10830.566 \end{bmatrix}$$

where, for Σ meaning summation over the sample blocks $h = 1, 2, \dots, 16$

$$z_1 = \Sigma A_h \bar{s}_{1h} / \Sigma A_h = 110570 / 4901 = 225.58049$$

= two stage sampling estimate of the average number of trees per acre of the entire forest area

$$z_2 = \Sigma A_h \bar{s}_{2h} / \Sigma A_h = 6812966.8 / 4901 = 1390.1177$$

= two stage sampling estimate of the average sum of tree diameters per acre of the entire forest area, and

$$z_3 = \Sigma A_h \bar{s}_{3h} / \Sigma A_h = 53080603 / 4901 = 10830.566$$

= two stage sampling estimate of the average sum of squared tree diameters per acre of the entire forest area

The ij -th element of the covariance matrix $[S]$ of $[z]$ can be calculated by the formula, say Main Formula

$$S_{z_i z_j} = \left(\frac{M-m}{M} \right) \left(\frac{m}{m-1} \right) \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (\Sigma A_h)^2 + \left(\frac{m}{M} \right) \Sigma \left(\frac{10A_h - n_h}{10A_h} \right) \left(\frac{A_{hh}^{2ij}}{n_h} \right) / (\Sigma A_h)^2$$

where Σ is taken over $h = 1, 2, \dots, 16$ sample clusters, and $N_h = 10A_h$. This requires the calculation of 16 covariance matrices of the variables s_1 , s_2 , and s_3 within blocks, say $[S_{hh}^{ss}]$, with its ij -th elements defined by the formula

$$\Sigma (s_{ihg} - \bar{s}_{ih}) (s_{jhg} - \bar{s}_{jh}) / (n_h - 1)$$

for $i, j = 1, 2, 3$, where Σ is taken over the n_h sample values s_{ihg} and s_{jhg} of sample plots $g = 1, 2, \dots, n_h$ within sample block h . The ij -th element of the covariance matrix of $[z]$ can also be calculated by any of the following alternative formulae.

Short-cut Formula 1

$$S_{z_i z_j} = m \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (m-1) (\Sigma A_h)^2 + m \Sigma A_h^2 S_{hh}^{ss} / n_h M (\Sigma A_h)^2$$

Short-cut Formula 2

$$S_{z_i z_j} = m \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (m-1) (\Sigma A_h)^2, \text{ and}$$

Short-cut Formula 3

$$S_{z_i z_j} = \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (\Sigma A_h)^2$$

It is more convenient, however, to define the covariance matrix $[S_{zz}]$ directly by matrix operations. For this we define first the 16 by 3 matrix of weighted differences

$$[V] = \begin{bmatrix} A_1(\bar{s}_{11}-z_1) & A_1(\bar{s}_{21}-z_2) & A_1(\bar{s}_{31}-z_3) \\ A_2(\bar{s}_{12}-z_1) & A_2(\bar{s}_{22}-z_2) & A_2(\bar{s}_{32}-z_3) \\ \vdots & \vdots & \vdots \\ A_{16}(\bar{s}_{1,16}-z_1) & A_{16}(\bar{s}_{2,16}-z_2) & A_{16}(\bar{s}_{3,16}-z_3) \end{bmatrix}$$

$$= \begin{bmatrix} -4147.1529 & -41048.644 & -578244.24 \\ -7523.9750 & -64677.634 & -720612.42 \\ \vdots & \vdots & \vdots \\ 12723.808 & 90200.423 & 744909.67 \end{bmatrix}$$

and then, the weighted sum of 16 matrices $[S_{ss}^{hh}]$, say

$$[S_{ss}] = \Sigma \left(\frac{10A_h - n_h}{10A_h} \right) \left(\frac{A_h^2}{n_h} \right) [S_{ss}^{hh}]$$

$$= \begin{bmatrix} 2012834975 & 10315563210 & 54028872200 \\ 10315563210 & 60558424930 & 424677708600 \\ 54028872200 & 424677708600 & 4484694027000 \end{bmatrix}$$

or the corresponding matrix for the short-cut formulae

$$[S_{ss}^*] = \Sigma (A_h^2/n_h) [S_{ss}^{hh}]$$

$$= \begin{bmatrix} 2022524766 & 10365134370 & 54287700110 \\ 10365134370 & 60849065330 & 426712725100 \\ 54287700110 & 426712725100 & 4506190215000 \end{bmatrix}$$

Using the Main Formula above, the covariance matrix $[S_{zz}]$ is calculated as

$$[S_{zz}] = \left(\frac{M-n}{M} \right) \left(\frac{m}{m-1} \right) [V]'[V]/(\Sigma A_h)^2 + \left(\frac{m}{M} \right) [S_{ss}^*]/(\Sigma A_h)^2$$

$$= \begin{bmatrix} 75.342297 & 365.40810 & 1591.8295 \\ 365.40810 & 2687.0073 & 25799.155 \\ 1591.8295 & 25799.155 & 397813.92 \end{bmatrix}$$

Consequently,

$$w = [b]'[z] = 106179.37 \text{ pounds}$$

= two-stage cluster sampling estimate of μ , the average biomass per acre for the entire forest area

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

$$= 51805483 + 11063125 = 62868608$$

= estimate of the variance of w

$$\sqrt{S_{ww}} = 7928.9727$$

= estimate of the standard error of w

and using a value $t = 2$,

$$w \pm t \sqrt{S_{ww}} = (106179 \pm 15858) \text{ pounds}$$

= 95 percent confidence limits of μ

Using now Short-cut Formula 1 we obtain the values

$$[S_{zz}] = m[V]'[V]/(m-1)(\Sigma A_h)^2 + m[S_{ss}^*]/M(\Sigma A_h)^2$$

$$= \begin{bmatrix} 83.874242 & 406.46033 & 1765.2882 \\ 406.46033 & 2997.9220 & 28877.038 \\ 1765.2882 & 28877.038 & 446493.73 \end{bmatrix}$$

$$S_{ww} = 58183864 + 11063125 = 69246988$$

and

$$w \pm t \sqrt{S_{ww}} = (106179 \pm 16643) \text{ pounds}$$

for an overestimation of the confidence interval of

$$(100)(16643 - 15858)/(15858) = 4.95 \text{ percent}$$

Using Short-cut Formula 2, we obtain

$$[S_{zz}] = m[V]'[V]/(m-1)(\Sigma A_h)^2$$

$$= \begin{bmatrix} 74.251111 & 357.14324 & 1506.9885 \\ 357.14324 & 2708.4035 & 26846.747 \\ 1506.9885 & 26846.747 & 425053.37 \end{bmatrix}$$

$$S_{ww} = 55707438 + 11063125 = 66770563$$

and

$$w \pm t \sqrt{S_{ww}} = (106179 \pm 16343) \text{ pounds}$$

for an overestimation of the confidence interval of

$$(100)(16343 - 15858)/(15858) = 3.06 \text{ percent}$$

Finally, using Short-cut Formula 3, we obtain

$$[S_{zz}] = [V]'[V]/(\Sigma A_h)^2$$

$$= \begin{bmatrix} 69.610416 & 334.82178 & 1412.8017 \\ 334.82178 & 2539.1282 & 25168.826 \\ 1412.8017 & 25168.826 & 398487.54 \end{bmatrix}$$

$$S_{ww} = 52225723 + 11063125 = 63288848$$

and

$$w \pm t \sqrt{S_{ww}} = (106179 \pm 15911) \text{ pounds}$$

for an overestimation of only

$$(100)(15911 - 15858)/(15858) = .33 \text{ percent}$$

We shall now consider the common procedure of calculating (1) the biomass of each tree by the regression function, (2) the biomass of the plot hk (k -th plot of h -th sample block) by summing up the biomass of all trees within plots, (3) the biomass per acre of plot hk , say y_{hk} by dividing the plot biomass by the plot area (one-tenth of an acre), (4) the average biomass per acre for each sample block $h = 1, 2, \dots, m$, say

$\bar{y}_h = \Sigma y_{hk}/n_h$, where Σ means summation over plot $k = 1, 2, \dots, n_h$ within block h and finally, (5) the estimate w of μ and the estimate of the variance of w by the formulae

$$w = \Sigma A_h \bar{y}_h / \Sigma A_h$$

and

$$S_{ww} = \left(\frac{M-m}{M} \right) \left(\frac{m}{m-1} \right) \Sigma A_h^2 (\bar{y}_h - w)^2 / (\Sigma A_h)^2 + \left(\frac{m}{M} \right) \Sigma \left(\frac{10A_h - n_h}{10A_h} \right) \left(\frac{A_h^2 S_{yy}^{hh}}{n_h} \right) / (\Sigma A_h)^2$$

where

$$S_{yy}^{hh} = \Sigma (y_{hk} - \bar{y}_h)^2 / (n_h - 1)$$

Of course, Σ of S_{ww} means summation over $h = 1, 2, \dots, 16$ sample blocks and Σ of S_{yy}^{hh} means summation over plots $k = 1, 2, \dots, n_h$ within block h .

Because the hk -th plot values s_{1hk} , s_{2hk} , and s_{3hk} have already been calculated, there is no need to calculate the individual tree biomass. We simply calculate the biomass per acre y_{hk} of plot hk by the formula

$$y_{hk} = b_1 s_{1hk} + b_2 s_{2hk} + b_3 s_{3hk} = [b]' [s_{hk}]$$

These values are listed by Cunia (1986d, Tables 7,8) and are not given here. Taking into account the classification of these plots into the 16 blocks, we continue with the calculation of the block average biomass per acre $\bar{y}_h$ and the sample variance within block S_{yy}^{hh} by the formulae above. These are listed in Table 1.

Finally, using these block values of means and variances, the reader can verify that

$$w = (\Sigma A_h \bar{y}_h) / (\Sigma A_h) = 520385103/4901 = 106179.37$$

and, by the formula above,

$$S_{ww} = 49340874 + 2464609 = 51805483$$

As the reader can verify, the estimate w of the average biomass per acre remains the same, while the estimate of the variance of w is equal only to the error component due to sample plots; the error component due to the biomass regression has been ignored. The variance is, thus, underestimated by

$$(100) (11063125) / (62868608) = 17.60 \text{ percent}$$

or in terms of standard errors

$$(100) (\sqrt{62868608} - \sqrt{51805483}) / (\sqrt{62868608}) = 9.22 \text{ percent}$$

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

- Cochran, W.G. Sampling Techniques, 3rd. Ed. John Wiley and Sons, New York, NY; 1977.
- Cunia, T. Some theory on the reliability of volume estimates in a forest inventory sample. Forest Science, 11: 115-128; 1965.
- Cunia, T. Error of the forest inventory estimates: its main components. In: Proceedings of the workshop on "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; 1986a.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "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; 1986b.
- Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the workshop on "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; 1986c.
- Cunia, T. On the error of biomass estimates in forest inventories: Part2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (86-001), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

ON THE ERROR OF FOREST INVENTORY ESTIMATES:

DOUBLE SAMPLING WITH REGRESSION^{1/}

Tiberius Cunia

Professor of Statistics and Operations Research,
SUNY College of Environmental Science and Forestry,
Syracuse, NY 13210

The error of biomass estimates in forest inventory contains a component due to the sample plots and another component due to the biomass regressions. A method is shown to estimate the error of the sample plots when they are selected by a double or two-phase sampling design. The error of the biomass regression function is assumed given.

Introduction

Most of the sampling designs for forest inventory consist of a relatively large sample of trees selected in clusters defined as trees growing within plots of fixed area or counted by a relascope at a Bitterlich sample point. These trees are measured for diameter, species and possibly other attributes other than biomass; and their biomass is estimated by means of previously determined regression functions. Using the estimated tree biomass one can then calculate estimates of the average biomass per unit area (by tree or stand classes). When the error of these estimates is calculated, however, the error of the biomass regressions is commonly ignored; only the error of the sample plots (or points) is taken into account.

An approach to combine the error of the biomass regression with that of the sample plots (or points) was proposed by Cunia (1965, 1986a). This approach requires that the true biomass regression function be of the form

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

where y is the tree biomass, $[x]$ is the vector of known tree attributes other than biomass and $[\beta]$ is the vector of unknown regression coefficients. Note that $[\]$ and $[\]'$ denote vectors or matrices and their transposes respectively. The vector $[\beta]$ is estimated from a sample of trees measured for biomass y and variables $[x]$ and we shall use the notation $[b]$ and $[S_{bb}]$ to denote the estimator of $[\beta]$ and covariance matrix of $[b]$ respectively.

^{1/}Paper based on a set of lecture notes: "On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots". Faculty of Forestry Miscellaneous Publication Number 9 (86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

The approach also requires that the parameter μ of interest be expressed as the product of two vectors, that is,

$$\mu = [\beta]'[z] = \mu_1 z_1 + \mu_2 z_2 + \dots + \mu_m z_m$$

where μ_{z_i} is the expected value (mean) of a function of the variable x_i , $i = 1, 2, \dots, m$. If μ is the average biomass per acre, then μ_{z_i} represents the expected values of x_i expressed on a per acre basis. For example, if x_i is the tree diameter d , the μ_{z_i} is (i) the average tree diameter, when μ is defined as the mean biomass per tree or (ii) average sum of tree diameters per acre, when μ is defined as mean biomass per acre. Similarly, when $x_i = 1$, then $\mu_{z_i} = 1$ or $\mu_{z_i} = \text{mean number of trees per acre depending on whether } \mu \text{ is the mean biomass per tree or per acre respectively.}$

It is assumed that the sample plots (or points) provide estimates z_i of μ_{z_i} and estimates $S_{z_i z_j}$ of the covariance of z_i and z_j , for $i, j = 1, 2, \dots, m$. Using the matrix notation, we define the vector $[z]$, the estimate of $[\mu_z]$ and the matrix $[S_{zz}]$, the estimate of the covariance matrix of $[z]$. Then, for the case where $[b]$ and $[z]$ are statistically independent, Cunia (1965, 1986a) has shown that

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

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z] \\ = \text{estimator of the variance of } w$$

Note how the variance of w can be viewed as having two additive components; one due to the error of the sample plots, the other due to the biomass regression.

The definition of the vector $[z]$ depends on (i) the sampling design by which the sample plots (or points) are selected, (ii) the parameter μ one wishes to estimate and (iii) the variables x_i used in the biomass regression. In his paper, Cunia (1986a) assumes that (i) the sample plots (or points) are selected by simple random sampling (or by a systematic sampling procedure that is equivalent to simple random sampling) and (ii) the parameter μ to estimate is the average biomass per plot. If the regression function used to calculate the tree biomass is $\hat{y} = [b]'[x]$ then, the estimators z_i are statistics based on the plot (or point) variables s_i defined as $(\sum x_i)/a$ for sample plots (or $\sum(x_i/a_k)$ for sample points) where a is the plot area in acres (or the factor to convert the variable x_i of the tree k of the sample point to a per acre basis). More specifically, for $n = \text{number of sample plots (or points),}$

$$z_i = \bar{s}_i = \sum s_i / n$$

and

$$S_{z_i z_j} = S_{s_i s_j} / n \\ = \sum (s_i - \bar{s}_i)(s_j - \bar{s}_j) / (n-1)$$

The objective of this paper is to extend the Cunia (1986a) approach from the simple random

sampling method by which the plots are selected to the method of double or two-phase sampling. More specifically we shall show how to calculate estimators $[z]$ and $[S_{zz}]$ of $[u_z]$ and $[s_{zz}]$ respectively by double sampling. We shall still be concerned with the estimator w of the mean biomass per acre u . We shall assume that we are given the estimates $[b]$ and $[S_{bb}]$ of the biomass regression function. For some of the methodology to calculate $[b]$ and $[S_{bb}]$ the interested reader is referred to Cunia (1986b,c) among others.

We shall also assume that the reader is familiar with the theory of double or two-phase sampling as described in standard texts on sampling techniques as that by Cochran (1977), for example. Before proceeding further with the procedures described in this paper, the reader is strongly advised to become familiar with the approach of taking into account the error of biomass regressions as described by Cunia (1986a). In particular he should understand very well the definition of the plot (or point) variables s , since this definition will be assumed known when the statistics z and their variances and covariances are defined. To simplify the discussion we shall use the terminology "plots" to denote both plots of fixed area and sample points; once the variables s are calculated, it does not matter whether the sample units are plots or points.

Double Sampling with Regression Applied to Forest Inventory

Sometimes, there is an auxiliary plot variable, say v , highly correlated with the plot biomass, say u . The variable v is easy to measure, but it is of little interest; and it is linearly related to the variable of interest u , that is relatively difficult to measure. Then, to estimate μ , the mean biomass per acre, one can use, with high efficiency, a double sampling with regression estimator. For example, let us assume that good, large scale aerial photography of a given forest area of interest exists. After locating a plot on an aerial photograph, we shall also assume that, by photogrammetry and photo-interpretation of the trees of the plot, one is able to derive an estimate v of the biomass per acre, that is highly correlated with the true biomass per acre u of the same plot as measured on the ground.

Consider now the following two-phase or double sampling with regression estimator design. In the first phase, n^* photo-plots are selected by simple random sampling without replacement, their values v are determined by photogrammetry and photo-interpretation and the statistics calculated from this sample are the average $\bar{v}^*$ and variance S_{vv}^* . In the second phase, n photo-plots are selected from the n^* plots of the first phase, again by simple random sampling without replacement. The plots are visited on the ground and measured for their true values u of biomass per acre. The sample data of the n pairs of values u and v provide the sample averages $\bar{u}$ and $\bar{v}$, variances and covariance S_{uu} , S_{vv} , and S_{uv} , and the linear regression and correlation coefficients

$c = S_{uv}/S_{vv}$ and $r = S_{uv}/\sqrt{S_{uu}S_{vv}}$ respectively. Then, the double sampling with regression estimator of the mean biomass per acre μ , and the estimator of its variance are defined respectively as

$$\bar{y}_r = \bar{u} - c(\bar{v} - \bar{v}^*)$$

and

$$S_{\bar{y}_r \bar{y}_r} = \left(\frac{n^* - n}{n^*} \right) S_{uu|v} \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + \left(\frac{N - n^*}{N} \right) \left(\frac{S_{uu}}{n^*} \right)$$

where N is the size of the finite population (possible number of photo-plots) and $S_{uu|v}$ is the estimator of the conditional variance of u given v calculated by the following formula, where Σ means summation over the n plots of the second phase subsample,

$$S_{uu|v} = \Sigma (u_h - u'_h)^2 / (n-2)$$

$$= (n-1)S_{uu}(1-r^2) / (n-2)$$

where

$$u'_h = (\bar{u} - c\bar{v}) + cv_h$$

= regression estimate of u_h of plot h
when $v = v_h$

Ordinarily, the population size N is infinite in size or very large compared to the first phase sample size n^* . Then, the finite population correction factor $(N - n^*)/N$ can be made equal to 1. We shall assume here that this is always the case and, thus, unless stated otherwise, we shall make $(N - n^*)/N = 1$.

Sometimes the first phase sample size n^* may be large with respect to the second phase subsample size n and, thus, $(n^* - n)/n$ can be made equal to 1. Then, the variance formula becomes

$$S_{\bar{y}_r \bar{y}_r} = S_{uu|v} \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + \frac{S_{uu}}{n^*}$$

When n is sufficiently large, the difference between $\bar{v}$ and $\bar{v}^*$ may become sufficiently small so that the effect of the term $(\bar{v} - \bar{v}^*)^2 / (n-1)S_{vv}$ may become negligibly small. Then, the variance formula changes to

$$S_{\bar{y}_r \bar{y}_r} = (n^* - n)S_{uu|v} / nn^* + S_{uu} / n^*$$

$$= S_{uu|v} / n - S_{uu|v} / n^* + S_{uu} / n^*$$

$$= S_{uu|v} / n + (S_{uu} - S_{uu|v}) / n^*$$

Note that in this last formula we have not made $(n^* - n)/n = 1$. Because, for the least squares linear regression, we have

$$S_{uu|v} = (n-1)(1-r^2)S_{uu} / (n-2) \approx (1-r^2)S_{uu}$$

we can further write

$$S_{\bar{y}_r \bar{y}_r} = S_{uu|v} / n + r^2 S_{uu} / n^*$$

Finally, when both n and n^* are sufficiently large and such that $(n^* - n)/n^*$ can be made equal to 1 and the effect of the factors $(\bar{v} - \bar{v}^*)^2 / (n-1)S_{vv}$ and $1/n^*$ is negligibly small, we can write

$$S_{\bar{y}_r \bar{y}_r} \approx S_{uu|v/n} + S_{uu/n^*} \approx S_{uu|v/n}$$

Seldom if ever one can determine the true value u of the ground plots. The usual procedure is to use a biomass regression function, estimate the biomass y of each tree in the plot by the regression value $\hat{y}$ and then calculate an estimate $\hat{u}$ of the true biomass u of the plot by adding the values $\hat{y}$ of all the trees in that plot. By using $\hat{u}$ instead of u in the formulae above results in an additional source of error (due to the biomass regression function) which is not accounted for when the variance of $\bar{y}_r$ is calculated. To take it into account, we shall use the following procedure.

We start with the assumption that u , the true value, is not known but can be estimated by the regression value

$$\hat{u} = b_1 s_1 + b_2 s_2 + \dots + b_m s_m = [b]'[s]$$

In this formula, the biomass regression function is

$$\hat{y} = b_1 x_1 + b_2 x_2 + \dots + b_m x_m = [b]'[x]$$

and the known estimator of the covariance matrix of $[b]$ is $[S_{bb}]$. For example, if the regression function is

$$\begin{aligned} \hat{y} &= b_1 + b_2 d + b_3 d^2 \\ &= b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x] \end{aligned}$$

where d = tree diameter, then, the plot variables s_1 , s_2 and s_3 are the usual plot values of "number of trees per acre", "sum of tree diameters per acre" and "sum of squared tree diameters per acre" respectively.

The estimator of $\bar{y}_r$ (the double sampling with regression estimator when the true values u_h are used), and through $\bar{y}_r$, the estimator of the parameter of interest μ (the mean biomass per acre) can be defined as

$$w = \bar{u} - c(\bar{v} - \bar{v}^*) = \bar{u} - S_{uv}(\bar{v} - \bar{v}^*)/S_{vv}$$

where

$$\bar{u} = \sum u_h/n = b_1 \bar{s}_1 + b_2 \bar{s}_2 + \dots + b_m \bar{s}_m$$

$$= [b]'[\bar{s}] = \sum [b]'[s]/n$$

and Σ is taken over the n sample plots h of the second phase.

It can be shown that

$$S_{uv} = b_1 S_{s_1 v} + b_2 S_{s_2 v} + \dots + b_m S_{s_m v} = [b]'[S_{sv}]$$

where

$$[S_{sv}]' = [S_{s_1 v} \quad S_{s_2 v} \quad \dots \quad S_{s_m v}]$$

Consequently, w can be successively written as

$$w = [b]'[\bar{s}] - [b]'[S_{sv}](\bar{v} - \bar{v}^*)/S_{vv}$$

$$= [b]'([\bar{s}] - [S_{sv}](\bar{v} - \bar{v}^*)/S_{vv})$$

$$= [b]'[z] = b_1 z_1 + b_2 z_2 + \dots + b_m z_m$$

where

$$z_1 = \bar{s}_1 - (S_{s_1 v}/S_{vv})(\bar{v} - \bar{v}^*) = \bar{s}_1 - c_1(\bar{v} - \bar{v}^*)$$

= double sampling with regression estimator of the mean of s_1

$$z_2 = \bar{s}_2 - (S_{s_2 v}/S_{vv})(\bar{v} - \bar{v}^*) = \bar{s}_2 - c_2(\bar{v} - \bar{v}^*)$$

= double sampling with regression estimator of the mean of s_2

$$z_3 = \bar{s}_3 - (S_{s_3 v}/S_{vv})(\bar{v} - \bar{v}^*) = \bar{s}_3 - c_3(\bar{v} - \bar{v}^*)$$

= double sampling with regression estimator of the mean of s_3

etc.

Of course, in these formulae, we have

$$c_1 = S_{s_1 v}/S_{vv}$$

= estimator of the linear regression coefficient of s_1 on v

$$c_2 = S_{s_2 v}/S_{vv}$$

= estimator of the linear regression coefficient of s_2 on v

$$c_3 = S_{s_3 v}/S_{vv}$$

= estimator of the linear regression coefficient of s_3 on v

etc.

Because $z_1, z_2, \dots, z_m$ are double sampling with regression estimators, their variances can be estimated by the formulae

$$\begin{aligned} S_{z_i z_i} &= \left(\frac{n^* - n}{n^*} \right) S_{s_i s_i} | v \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + S_{s_i s_i}/n^* \\ &= \text{estimator of the variance of } z_i, \text{ for } i = 1, 2, \dots, m \end{aligned}$$

where

$$\begin{aligned} S_{s_i s_i} | v &= \Sigma (s_{hi} - s'_{hi})^2 / (n-2) \\ &= (n-1) S_{s_i s_i} (1 - r_i^2) / (n-2) \end{aligned}$$

$$\begin{aligned} s'_{hi} &= (\bar{s}_i - c_i \bar{v}) + c_i v_h \\ &= \text{regression estimate of } s_{hi} \text{ given that } v = v_h \end{aligned}$$

and

$$r_i = S_{s_i v} / \sqrt{S_{s_i s_i} S_{vv}}$$

= estimator of the linear correlation coefficient of s_i and v

Similarly, the covariance of z_i and z_j , for $i \neq j = 1, 2, \dots, m$ can be estimated by the formula

$$S_{z_i z_j} = \left(\frac{n^* - n}{n^*} \right) S_{s_i s_j} | v \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + S_{s_i s_j}/n^*$$

Where, for Σ meaning summation over clusters h ,

$$\begin{aligned} S_{s_i s_j} | v &= \Sigma (s_{hi} - s'_{hi})(s_{hj} - s'_{hj}) / (n-2) \\ &= (n-1) S_{s_i s_j} (1 - r_i r_j / r_{ij}) / (n-2) \end{aligned}$$

= estimator of the conditional variance of s_i and s_j when v is given,

$$\begin{aligned} s'_{hi} &= (\bar{s}_i - c_i \bar{v}) + c_i v_h \\ &= \text{regression estimate of } s_{hi} \text{ given that } v = v_h \end{aligned}$$

$$s'_{hj} = (\bar{s}_j - c_j \bar{v}) + c_j v_h$$

= regression estimate of s_{hj} given that $v = v_h$

$$r_i = S_{s_i v} / \sqrt{S_{s_i s_i} S_{v v}}$$

= estimator of the linear correlation coefficient of s_i and v

$$r_j = S_{s_j v} / \sqrt{S_{s_j s_j} S_{v v}}$$

= estimator of the linear correlation coefficient of s_j and v

and

$$r_{ij} = S_{s_i s_j} / \sqrt{S_{s_i s_i} S_{s_j s_j}}$$

= estimator of the linear correlation coefficient of s_i and s_j

We have shown above the shortcut, approximate formulae for the calculation of the variance of the double sampling with regression estimator $\bar{y}_r$. Similarly, for the variance of z_i , $i = 1, 2, \dots, m$ we can write

$$S_{z_i z_i} = S_{s_i s_i} |v \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{v v}} \right) + \frac{S_{s_i s_i}}{n^*}$$

when n^* is sufficiently large with respect to n

$$S_{z_i z_i} = (n^* - n) S_{s_i s_i} |v / nn^* + S_{s_i s_i} / n^*$$

$$= S_{s_i s_i} |v / n + r_i^2 S_{s_i s_i} / n^*$$

when n is sufficiently large, and

$$S_{z_i z_i} \approx S_{s_i s_i} |v / n + S_{s_i s_i} / n^* \approx S_{s_i s_i} |v / n$$

when both n and n^* are sufficiently large and such that $(n^* - n) / n^* \approx 1$ the effect of the factors $1/n^*$ and $(\bar{v} - \bar{v}^*)^2 / (n-1)S_{v v}$ is negligibly small.

Similarly, one can write the following shortcut approximate formulae for the covariance of z_i and z_j , $i \neq j = 1, 2, \dots, m$

$$S_{z_i z_j} = S_{s_i s_j} |v \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{v v}} \right) + \frac{S_{s_i s_j}}{n^*}$$

when n^* is sufficiently large with respect to n

$$S_{z_i z_j} = (n^* - n) S_{s_i s_j} |v / nn^* + S_{s_i s_j} / n^*$$

$$= S_{s_i s_j} |v / n + (r_i r_j / r_{ij}) S_{s_i s_j} / n^*$$

when n is sufficiently large, and

$$S_{z_i z_j} \approx S_{s_i s_j} |v / n + S_{s_i s_j} / n^* \approx S_{s_i s_j} |v / n$$

when both n and n^* are sufficiently large and such that $(n^* - n) / n^* \approx 1$ and the effect of the factors $(\bar{v} - \bar{v}^*)^2 / (n-1)S_{v v}$ and $1/n^*$ is negligibly small.

As all the elements of the covariance matrix $[S_{zz}]$ of $[z]$ are now defined, the formula for the variance of $w = [b]'[z]$ follows immediately as

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

An Illustrative Example

A forest area has been sampled for biomass in 1960; the trees from 235 randomly selected permanent sample plots were all measured for their diameter values d_1 in inches. Four years later, in 1964 it was desired to update the forest biomass estimates. For this purpose a subsample of 118 of the old 235 plots was selected at random and its trees were measured again for their new diameters d_2 . By defining the 1960, the first measurement plots as the first phase sample and the 1964, the second measurement plots as the second phase sample we can calculate estimators of μ_2 , the mean biomass per acre at the second, 1964 measurement time by the following, double sampling with regression estimator procedures.

Procedure 1: Ignore the error of the biomass regression function and use as auxiliary variable the plot value

v = estimate of the plot biomass per acre at the first, 1960 measurement time.

This is the common, double sampling with regression estimator procedure.

Procedure 2: Take the error of the biomass regression into account and use as auxiliary variable the plot value

v = estimate of the plot biomass per acre at the first, 1960 measurement time.

This is double sampling with (simple linear) regression estimator of the previous section.

Procedure 3: Take the error of the biomass regression into account and use as auxiliary variables the plot values $s_1, s_2, \dots, s_m$ of the first, 1960 measurement time. This is the double sampling with (multiple linear) regression estimator a simple extension of the double sampling with (simple linear) regression estimator of the previous section.

In estimating the tree and plot biomass regression we shall use the biomass regression function calculated by weighted least squares in Cunia (1986b) and already used in Example 1 of Cunia (1986a). For convenience, the statistics of this parabolic regression, that is,

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

$$= b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x]$$

where d = tree diameter (of first or second measurement) and the definitions of x_1, x_2 and x_3 are obvious, are shown below

$$[b]' = [5.1818118 \quad -25.653078 \quad 12.988357]$$

and

$$[S_{bb}] = \begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

The use of this regression function implies that the plot variables (that summarize the plot information) are

s_1 = number of trees per acre
 s_2 = sum of diameters per acre
 s_3 = sum of squared diameters per acre

There are three variables s measured in 1960 on all 235 sample plots and three additional variables s measured in 1964 on the 118 plots of the subsample. Using the first set of three variables s (the 1960 values) one can calculate the variable

$$v = b_1 s_1 + b_2 s_2 + b_3 s_3 = [b]'[s] \\ = 1960 \text{ biomass per acre}$$

and using the second set of three variables s (the 1964 values) one can calculate the variable

$$u = b_1 s_1 + b_2 s_2 + b_3 s_3 = [b]'[s] \\ = 1964 \text{ biomass per acre}$$

The plot values s_1 , s_2 , s_3 , v and u of the first and second measurement are listed by Cunia (1986d, Tables 1, 2, 3, 7, 8); they are not repeated here. We shall only report the summary values, as they are needed by each procedure.

Procedure 1: The estimator of μ_2 , the mean biomass per acre at the second, 1964 occasion, is the classical double sampling with regression estimator, where the error of the biomass regression function is being ignored, when the error of the estimator $\bar{y}_r$ of μ_2 is being calculated. The variable of interest is u = plot biomass per acre at the second measurement and the auxiliary variable is v = plot biomass per acre at the first measurement. The two variables are highly and positively correlated.

Let us now calculate the necessary statistics. We shall start with the basic sums, sums of squares and sums of cross-products. With Σ meaning summation over the values of the 235 plots measured in 1960 and listed by Cunia (1986d, Tables 7, 8) we find the following sum

$$\Sigma v = 24897913$$

For Σ meaning summation over the 118 permanent sample plot values v and u of the first (1960) and second (1964) measurement respectively listed by Cunia (1986d, Table 8) we find

$$\begin{aligned} \Sigma v &= 12490431, & \Sigma v^2 &= 1969954828000 \\ \Sigma u &= 14084491, & \Sigma u^2 &= 2295905587000 \\ & \text{and } \Sigma uv &= 2105734088000 \end{aligned}$$

Using the above sums and the notation of the previous section we calculate then

$$\begin{aligned} \bar{v}^* &= \Sigma v/n^* = 24897913/235 = 105948.56 \\ &= \text{estimate of the average biomass per acre at the first, 1960 measurement time; all 235 sample plot data were used.} \end{aligned}$$

$$\begin{aligned} \bar{v} &= \Sigma v/n = 12490431/118 = 105851.11 \\ &= \text{estimate of the average biomass per acre at the first, 1960 measurement time;} \end{aligned}$$

only the data from the 118 permanent plots measured on both occasions were used.

$$\begin{aligned} \bar{u} &= \Sigma u/n = 14084491/118 = 119360.09 \\ &= \text{estimate of the average biomass per acre at the second, 1964 measurement time; only the data from the 118 permanent plots measured on both occasions were used.} \end{aligned}$$

$$\begin{aligned} S_{vv} &= \Sigma (v - \bar{v})^2 / (n-1) = 5536998459 \\ &= \text{estimate of the variance of } v_h \text{ based on the 118 permanent plots of the second phase subsample only.} \end{aligned}$$

$$\begin{aligned} S_{uv} &= \Sigma (u - \bar{u})(v - \bar{v}) / (n-1) = 5255342828 \\ &= \text{estimate of the covariance of } u_h \text{ and } v_h \end{aligned}$$

$$\begin{aligned} S_{uu} &= \Sigma (u - \bar{u})^2 / (n-1) = 5254525850 \\ &= \text{estimate of the variance of } u_h \end{aligned}$$

$$\begin{aligned} c &= S_{uv}/S_{vv} = .94913207 \\ &= \text{estimate of the regression coefficient of the linear regression of } u \text{ on } v \end{aligned}$$

$$\begin{aligned} r^2 &= S_{uv}^2 / S_{vv} S_{uu} = .94927965 \\ &= \text{square of the estimate of the linear correlation coefficient of } u \text{ and } v, \text{ and} \end{aligned}$$

$$\begin{aligned} S_{uu|v} &= (n-1) S_{uu} (1-r^2) / (n-2) = 268808928 \\ &= \text{estimate of the (conditional) variance of } u \text{ about the least squares regression line of } u \text{ on } v \end{aligned}$$

We can now calculate the double sampling with (linear) regression estimator of μ_2 , the mean 1964 biomass per acre when (1) the auxiliary variable is the biomass per acre v at the first measurement time in 1960, and (2) the error in the biomass regression function is ignored.

$$\begin{aligned} \bar{y}_r &= \bar{u} - c(\bar{v} - \bar{v}^*) \\ &= 119360.09 - (.94913207)(-97.453529) \\ &= 119360.09 + 92.496270 = 119452.59 \\ &= \text{double sampling with regression estimate of } \mu_2, \text{ the average biomass per acre at the second, 1964 measurement time} \end{aligned}$$

$$\begin{aligned} S_{\bar{y}_r \bar{y}_r} &= \left(\frac{n^* - n}{n^*} \right) S_{uu|v} \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + \frac{S_{uu}}{n^*} \\ &= (.49787234)(268808928)(.0084745909) \\ &\quad + 22359684 \\ &= 23493860 \\ &= \text{estimate of the variance of } \bar{y}_r \end{aligned}$$

$$\sqrt{S_{\bar{y}_r \bar{y}_r}} = \sqrt{23493860} = 4847.0466$$

= estimate of the standard error of $\bar{y}_r$ and using $t = 2$

$$\begin{aligned} \bar{y}_r \pm t \sqrt{S_{\bar{y}_r \bar{y}_r}} &= (119453 \pm 9694) \text{ pounds} \\ &= 95 \text{ percent confidence limits of } \mu_2 \end{aligned}$$

It may be interesting to see the approximation given by various shortcut formulae.

(1) Although n^* is not large with respect to n , and the value of the correction factor $(n^*-n)/n^* = 117/235$ is approximately equal to .5, we shall use the corresponding variance formula and obtain

$$S_{\bar{y}_R \bar{y}_R} \approx S_{uu} | v \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + \frac{S_{uu}}{n^*} = 24637730$$

The 95 percent confidence limits become

$$(119453 \pm 9927) \text{ pounds}$$

As the reader can verify, the variance is overestimated by

$$\frac{(100)(24637730 - 23493860)}{23493860} = 4.87 \text{ percent}$$

while the 95 percent bound on the error of $\bar{y}_R$ is overestimated by

$$(100)(9927 - 9694)/(9694) = 2.40 \text{ percent}$$

This seems to be an acceptable approximation.

(2) The value of $n = 118$ is sufficiently large and the approximation given by the formula

$$S_{\bar{y}_R \bar{y}_R} \approx \left(\frac{n^*-n}{n^*} \right) S_{uu} | v/n + S_{uu}/n^* \\ = S_{uu} | v/n + r^2 S_{uu}/n^* = 23503635$$

is excellent; the variance of $\bar{y}_R$ is overestimated now by only .04 percent and the 95 percent confidence limits remain practically the same

$$(119453 \pm 9696) \text{ pounds}$$

(3) Using the formula where $(n^*-n)/n^* = 1$, $1/n^* = 0$ and $(\bar{v} - \bar{v}^*)^2/(n-1)S_{vv} = 0$, we have

$$S_{\bar{y}_R \bar{y}_R} = S_{uu} | v/n = 22780417$$

and the 95 percent confidence limits of

$$(119453 \pm 9546) \text{ pounds}$$

practically the same thing as in (2) above.

Procedure 2: To simplify notation, we shall use s_1 , s_2 , and s_3 to denote the variables of the second measurement only; the corresponding variables of the first measurement are only needed to calculate the auxiliary variable v .

By this procedure, the estimate of the average biomass per acre μ_2 , at the second, 1964 measurement time is defined as

$$w = [b]'[z]$$

where

$$z_1 = \bar{s}_1 - c_1(\bar{v} - \bar{v}^*) \\ = \text{double sampling with regression estimator of the population mean "number of trees per acre," say } \mu_{s_1}, \text{ at the second, 1964 measurement time}$$

$$z_2 = s_2 - c_2(\bar{v} - \bar{v}^*) \\ = \text{double sampling with regression estimator of the population mean "sum of tree diameters per acre," say } \mu_{s_2}, \text{ at the second, 1964 measurement time}$$

$$z_3 = \bar{s}_3 - c_3(\bar{v} - \bar{v}^*)$$

= double sampling with regression estimator of the population mean "sum of squared tree diameters per acre," say μ_{s_3} , at the second, 1964 measurement time

To calculate z_1 , z_2 , and z_3 , we need the following statistics. For Σ meaning summation over plots $h = 118, 119, \dots, 235$, and using the individual plot values s_1 , s_2 , s_3 (of the second measurement) and v (of the first measurement) as listed by Cunia (1986d, Tables 2, 8) we calculate the following statistics

$$S_{s_1 s_1} = \Sigma(s_1 - \bar{s}_1)^2/(n-1) = 24039.555 \\ = \text{estimate of the variance of } s_1$$

$$S_{s_1 s_2} = \Sigma(s_1 - \bar{s}_1)(s_2 - \bar{s}_2)/(n-1) = 123557.78 \\ = \text{estimate of the covariance of } s_1 \text{ and } s_2$$

$$S_{s_1 s_3} = \Sigma(s_1 - \bar{s}_1)(s_3 - \bar{s}_3)/(n-1) = 635935.23 \\ = \text{estimate of the covariance of } s_1 \text{ and } s_3$$

$$S_{s_2 s_2} = \Sigma(s_2 - \bar{s}_2)^2/(n-1) = 718316.31 \\ = \text{estimate of the variance of } s_2$$

$$S_{s_2 s_3} = \Sigma(s_2 - \bar{s}_2)(s_3 - \bar{s}_3)/(n-1) = 4802637.7 \\ = \text{estimate of the covariance of } s_2 \text{ and } s_3$$

$$S_{s_3 s_3} = \Sigma(s_3 - \bar{s}_3)^2/(n-1) = 47000205 \\ = \text{estimate of the variance of } s_3$$

$$S_{vv} = \Sigma(v - \bar{v})^2/(n-1) = 55369985 \\ = \text{estimate of the variance of } v$$

$$S_{s_1 v} = \Sigma(s_1 - \bar{s}_1)(v - \bar{v})/(n-1) = 4357434.1 \\ = \text{estimate of the covariance of } s_1 \text{ and } v$$

$$S_{s_2 v} = \Sigma(s_2 - \bar{s}_2)(v - \bar{v})/(n-1) = 41121209 \\ = \text{estimate of the covariance of } s_2 \text{ and } v$$

$$S_{s_3 v} = \Sigma(s_3 - \bar{s}_3)(v - \bar{v})/(n-1) = 484098884 \\ = \text{estimate of the covariance of } s_3 \text{ and } v$$

$$c_1 = S_{s_1 v}/S_{vv} = .00078696683 \\ = \text{estimate of the linear regression coefficient of } s_1 \text{ on } v$$

$$c_2 = S_{s_2 v}/S_{vv} = .0074266246 \\ = \text{estimate of the linear regression coefficient of } s_2 \text{ on } v$$

$$c_3 = S_{s_3 v}/S_{vv} = .087429839 \\ = \text{estimate of the linear regression coefficient of } s_3 \text{ on } v$$

This yields immediately the values

$$z_1 = 266.01695 \\ - (.00078696683)(105851.11 - 105948.56) \\ = 266.01695 + .076692694 = 266.09364 \\ = \text{double sampling with regression estimate of } \mu_{s_1}, \text{ the mean "number of trees per acre" at the second, 1964 measurement time}$$

$$\begin{aligned}
z_2 &= 1638.6805 \\
&- (.0074266246)(105851.11 - 105948.56) \\
&= 1638.6805 + .72375078 = 1639.4043 \\
&= \text{double sampling with regression estimate of } \mu_{s_2}, \text{ the mean "sum of tree diameters per acre" at the second, 1964 measurement time}
\end{aligned}$$

and

$$\begin{aligned}
z_3 &= 12320.176 \\
&- (.087429839)(105851.11 - 105948.56) \\
&= 12320.176 + 8.5203463 = 12328.696 \\
&= \text{double sampling with regression estimate of } \mu_{s_3}, \text{ the mean "sum of squared diameters per acre" at the second, 1964 measurement time}
\end{aligned}$$

The variances and covariances of z_1 , z_2 , and z_3 can now be estimated by the formulae

$$S_{z_i z_j} = \left(\frac{n^* - n}{n^*} \right) S_{s_i s_j | v} \left(\frac{1}{n} - \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + \frac{S_{s_i s_j}}{n^*}$$

for $i, j = 1, 2, 3$

where, for Σ meaning summation over plots $h = 118, 119, \dots, 235$,

$$\begin{aligned}
S_{s_i s_j | v} &= \Sigma (s_i - (\bar{s}_i - c_i \bar{v}) - c_i v) (s_j - (\bar{s}_j - c_j \bar{v}) - c_j v) / (n-2) \\
&= (n-1) S_{s_i s_j} (1 - r_i r_j / r_{ij}) / (n-2)
\end{aligned}$$

$$\begin{aligned}
r_i &= S_{s_i v} / \sqrt{S_{s_i s_i} S_{vv}} \\
&= \text{estimate of the linear correlation coefficient of } s_i \text{ and } v
\end{aligned}$$

and

$$\begin{aligned}
r_{ij} &= S_{s_i s_j} / \sqrt{S_{s_i s_i} S_{s_j s_j}} \\
&= \text{estimate of the linear correlation coefficient of } s_i \text{ and } s_j
\end{aligned}$$

Obviously, for any $i = 1, 2, 3$, we have $r_{ii} = 1$

Doing the necessary algebraic calculations we find first

$$\begin{aligned}
S_{s_1 s_1 | v} &= 20788.075 & S_{s_1 s_2 | v} &= 91982.936 \\
S_{s_1 s_3 | v} &= 257163.45 & S_{s_2 s_2 | v} &= 416484.22 \\
S_{s_2 s_3 | v} &= 1217825.8 \text{ and } S_{s_3 s_3 | v} &= 4715823.6
\end{aligned}$$

Then, we calculate the covariances of z_i and z_j and arrange them in the following covariance matrix

$$[S_{zz}] = \begin{bmatrix} 102.29583 & 525.77714 & 2706.1055 \\ 525.77714 & 3056.6621 & 20436.747 \\ 2706.1055 & 20436.747 & 200000.84 \end{bmatrix}$$

Consequently,

$$\begin{aligned}
w &= [b]'[z] = 119452.59 \\
&= \text{double sampling with regression estimate of } \mu_2, \text{ the average biomass per acre at the second, 1964 measurement time}
\end{aligned}$$

$$\begin{aligned}
S_{ww} &= [b]'[S_{zz}][b] + [z]'[S_{bb}][z] \\
&= 22359683 + 11407373 = 33767056 \\
&= \text{estimate of the variance of } w
\end{aligned}$$

$$\begin{aligned}
\sqrt{S_{ww}} &= \sqrt{33767056} = 5810.9427 \\
&= \text{estimate of the standard error of } w
\end{aligned}$$

and using a t-value of 2,

$$\begin{aligned}
w \pm t\sqrt{S_{ww}} &= (119453 \pm 11622) \text{ pounds} \\
&= 95 \text{ percent confidence limits of } \mu_2
\end{aligned}$$

As the reader can verify, ignoring the effect of the error of biomass tables, results in an underestimation of the variance of w by

$$(100)(33767056 - 23493860) / (33767056) = 30.42 \text{ percent}$$

The standard error of w is underestimated by

$$(100)(5810.9427 - 4847.0466) / (5810.9427) = 16.59 \text{ percent}$$

While the estimate of μ_2 is exactly the same by the two procedures, the estimate of the variance of w by procedure 1 is slightly higher than the value of the first component of the variance as estimated by the procedure 2. The difference is small, only

$$(100)(23493860 - 22359683) / (23493860) = 4.83 \text{ percent}$$

Procedure 3: It is felt that the regression estimators of the mean of s_1 , s_2 and s_3 (of the second measurement time) of Procedure 2 above may be improved if the auxiliary variable v is replaced by the first measurement values of the variables s_1 , s_2 and s_3 . Recall that v is a linear combination of the first measurement values of s_1 , s_2 and s_3 . To simplify notation we shall use

- (1) s_1 , s_2 and s_3 to denote the variables of the second measurement and
- (2) $v_1 = s_1$ of the first measurement
 $v_2 = s_2$ of the first measurement, and
 $v_3 = s_3$ of the first measurement.

The values v_1 , v_2 and v_3 are to be found in Cunia (1986d, Tables 1, 2) and the values s_1 , s_2 and s_3 in Cunia (1986d, Table 2). Using the first measurement values $v_1 = s_1$, $v_2 = s_2$ and $v_3 = s_3$ of the 235 plots that were measured in 1964 we find

$$[\bar{v}^*] = \begin{bmatrix} \bar{v}_1^* \\ \bar{v}_2^* \\ \bar{v}_3^* \end{bmatrix} = \begin{bmatrix} 225.61702 \\ 1389.0374 \\ 10810.647 \end{bmatrix}$$

= vector of the average values of the variables v_1 , v_2 and v_3 as calculated from the first measurement data of the entire sample of 235 plots

Using now the first and the second measurement values of the subsample of 118 permanent plots measured in 1960 and 1964 (variables v_1 , v_2 and v_3 for the first, and variables s_1 , s_2 and s_3 for the second measurement) we find the following statistics

$$[\bar{v}] = \begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \bar{v}_3 \end{bmatrix} = \begin{bmatrix} 236.35593 \\ 1439.3161 \\ 10898.164 \end{bmatrix}$$

= vector of the sample averages of v_1 , v_2 and v_3 of the 118 permanent plots

$$[S_{vv}] = \begin{bmatrix} 21272.932 & 110452.37 & 602351.30 \\ 110452.37 & 659011.32 & 4724541.6 \\ 602351.30 & 4724541.6 & 48604076 \end{bmatrix}$$

= estimate of the covariance matrix of v_1 , v_2 and v_3 as calculated from the data of the 118 permanent plots

$$[\bar{s}] = \begin{bmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \bar{s}_3 \end{bmatrix} = \begin{bmatrix} 266.01695 \\ 1638.6805 \\ 12320.176 \end{bmatrix}$$

= vector of the sample averages of s_1 , s_2 and s_3 of the 118 permanent plots

$$[S_{ss}] = \begin{bmatrix} 24039.555 & 123557.78 & 635935.23 \\ 123557.78 & 718316.31 & 4802637.7 \\ 635935.23 & 4802637.7 & 47000205 \end{bmatrix}$$

= estimate of the covariance matrix of s_1 , s_2 and s_3 as calculated from the data of the 118 permanent plots

$$[S_{sv}] = \begin{bmatrix} 21987.926 & 109292.82 & 542577.80 \\ 118295.97 & 670898.34 & 4443890.0 \\ 659678.73 & 4832348.2 & 46552860 \end{bmatrix}$$

= estimate of the covariance matrix of s_1 , s_2 and s_3 with v_1 , v_2 and v_3 as calculated from the data of the 118 permanent plots

Note that the first row of $[S_{sv}]$ contains the covariances of s_1 with v_1 , v_2 , v_3 , the second column of $[S_{sv}]$ contains the covariances of v_2 with s_1 , s_2 and s_3 , etc.

Using the above values we calculate

$$[C] = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = [S_{vv}]^{-1} [S_{sv}]'$$

$$= \begin{bmatrix} 1.7649801 & 3.5411706 & 9.0746818 \\ -.17546979 & .27603189 & -.81992294 \\ .0063462595 & .020713020 & .92503505 \end{bmatrix}$$

= estimate of the matrix of regression coefficients of s_1 (c_{11} , c_{21} , c_{31}), s_2 (c_{12} , c_{22} , c_{32}) and s_3 (c_{13} , c_{23} , c_{33}) as the dependent variables and v_1 , v_2 and v_3 as the independent variables.

Note that the regression function of s_1 on v_1 , v_2 and v_3 can be written as

$$s_1' = \bar{s}_1 + c_{11}(v_1 - \bar{v}_1) + c_{21}(v_2 - \bar{v}_2) + c_{31}(v_3 - \bar{v}_3)$$

$$= 32.247354 + 1.7649801 v_1 - .17546979 v_2 + .0063462595 v_3$$

with similar expressions for the regression functions of s_2 and s_3 .

We can now calculate the vector of statistics $[z]$ by the formula

$$[z] = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = [\bar{s}] - [C]' [\bar{v} - \bar{v}'] = \begin{bmatrix} 255.32996 \\ 1584.9609 \\ 12182.992 \end{bmatrix}$$

To estimate the covariance matrix of $[z]$ we need to calculate first the matrix $[R]$ of the residuals of s_1 , s_2 and s_3 from their own regressions. The first column of $[R]$ contains the 118 residuals ($s_1 - s_1'$) defined as

$$s_1 - s_1' = (s_1 - \bar{s}_1) - c_{11}(v_1 - \bar{v}_1) - c_{21}(v_2 - \bar{v}_2) - c_{31}(v_3 - \bar{v}_3)$$

Defined by similar formulae, the two sets of 118 residuals ($s_2 - s_2'$) and ($s_3 - s_3'$) are contained in the second and third column respectively of $[R]$. Dividing the product $[R]'[R]$ by the number ($n-4$) = 114 degrees of freedom yields the covariance matrix

$$[S_{ss|v}] = [R]'[R]/114$$

$$= \begin{bmatrix} 990.96189 & 4400.9063 & 24744.489 \\ 4400.9063 & 22757.914 & 172903.71 \\ 24744.489 & 172903.71 & 1963296.7 \end{bmatrix}$$

= estimate of the conditional covariance matrix of the variables s_1 , s_2 and s_3 for given v_1 , v_2 and v_3 , calculated from the data of the 118 plots of the second phase subsample.

This yields the statistic

$$[S_{zz}] = \left(\frac{n^* - n}{n^*} \right) \left([\bar{v} - \bar{v}']' [S_{vv}]^{-1} [\bar{v} - \bar{v}'] \right) [S_{ss|v}] + \frac{1}{n^*} [S_{ss}]$$

$$= \begin{bmatrix} 108.87373 & 554.98988 & 2870.3549 \\ 554.98988 & 3207.7263 & 21584.446 \\ 2870.3549 & 21584.446 & 213032.73 \end{bmatrix}$$

= estimate of the covariance

The point and interval estimates of the average biomass per acre μ_2 at the second, 1964 measurement time follows immediately as

$$w = [b]'[z] = 118900.99 = \text{double sampling with (multiple linear) regression estimate of } \mu_2$$

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

$$= 23907253 + 11798043 = 35705297$$

= estimate of the variance of w

$$\sqrt{S_{ww}} = \sqrt{35705297} = 5975.3909$$

= estimate of the standard error of w

and using a t-value of 2,

$$w \pm 2\sqrt{S_{ww}} = (118901 \pm 11951) \text{ pounds}$$

= 95 percent confidence limits of μ_2

The reader may be surprised by the fact that the error of the estimate by Procedure 2 (using simple linear regression) is smaller than the error by Procedure 3 (that uses multiple linear regression). The standard error of w is 5811 by Procedure 2 and 5975 by Procedure 3, an increase of some 2.8 percent. This may be due to sampling error. A more credible reason is, however, that the multiple linear regression of s_1 on v_1 , v_2

and v_3 may not necessarily be better than the simple linear regression of s_i on their linear combination $v = b_1v_1 + b_2v_2 + b_3v_3$. The error of the multiple linear regression has four sources (the four regression coefficients) compared to the error of the simple linear regression that has only two (the two regression coefficients). Consequently, whenever the multiple linear regression is not significantly better, it may well be possible for the corresponding double sampling estimator to be less precise than the double sampling with simple linear regression estimator.

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

- Cochran, W. G. Sampling Techniques, 3rd Ed.
John Wiley and Sons, New York, NY; 1977.
- Cunia, T. Some theory on the reliability of volume estimates in a forest inventory sample. Forest Science, 11:115-128; 1965.
- Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the workshop on "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; 1986a.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "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; 1986b.
- Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the workshop on "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; 1986c.
- Cunia, T. On the error of biomass estimates in forest inventories: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (86-001), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

ON THE ERROR OF FOREST INVENTORY ESTIMATES:

CONTINUOUS FOREST INVENTORY WITHOUT SPR⁽¹⁾

Tiberius Cunia

Professor of Statistics and Operations Research
SUNY College of Environmental Science and Forestry
Syracuse, NY, 13210

The error of the biomass estimates in Continuous Forest Inventory Systems has two main error components; one due to sample plots and one due to biomass tables or regressions. The common procedures by which the estimates are calculated take into account only the first component; the second component is simply ignored. An approach is shown that introduces the error of the biomass regressions into the total error of the estimates of average biomass per tree, average biomass per acre and growth components such as average net change from one to the next occasion, average mortality or average ingrowth biomass per acre.

Introduction

The sampling design of a forest inventory system consists generally of a random selection of sample plots (or Bitterlich relascope points) where the trees are measured for diameter and where the biomass of these sample trees is estimated by biomass tables or regression functions. When inferences about the error of the forest biomass estimates are made, however, only the error from the sample plots (or points) is taken into account; the error of the biomass tables or regressions is simply ignored.

Cunia (1986a) has proposed an approach to combine the error of the biomass regressions with that of the sample plots (or points). This approach requires that the estimators be of the form

$$w = [b]'[z] = b_1 z_1 + b_2 z_2 + \dots + b_m z_m$$

where $[b]$ is the estimate of the vector of coefficients of the biomass regression

$$\hat{y} = [b]'[x] = b_1 x_1 + b_2 x_2 + \dots + b_m x_m$$

y is the tree biomass and $[z]$ is the vector of statistics calculated from the sample plot data. We implicitly assume here that the statistics z are defined so that the product of the expected value of the vector $[z]$, say $[\mu_z]$ with the expected value of the vector $[b]$, say $[\beta]$ is equal

to the parameter of interest μ , say $\mu = [\beta]'[\mu_z]$. The variance of w can be estimated by the approximate formula

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

where $[S_{zz}]$ and $[S_{bb}]$ are the estimates of the covariance matrices of $[z]$ and $[b]$ respectively. Note that the first part of S_{ww} may be viewed as the error component due to sample plots and the second part may be viewed as the error component due to biomass regressions.

The definition of $[z]$ depends on the (i) specific parameter μ one wishes to estimate, (ii) specific sampling design by which the sample plots are selected (including the type of sample units used) and (iii) specific independent variables $x_1, x_2, \dots, x_m$ used in the biomass regression function. In the Cunia (1986a) paper it is assumed that (i) the parameter to estimate is the average biomass per acre μ , (ii) the sample plots (or points) are selected by simple random sampling and (iii) the definition of the statistics z is based on the plot (or point) variables $s_1, s_2, \dots, s_m$ defined as the averages of the regression variables $x_1, x_2, \dots, x_m$ expressed on a "per acre" basis.

For example, let us assume that $x_1 = 1$, $x_2 = d$ = tree diameter and $x_3 = d^2$. Then, for Σ meaning summation over the trees of a given plot, we have the plot variables

$$\begin{aligned} s_1 &= (\Sigma x_1)/(\text{plot area}) \\ &= \text{number of trees per acre} \\ s_2 &= (\Sigma x_2)/(\text{plot area}) \\ &= \text{sum of tree diameters per acre, and} \\ s_3 &= (\Sigma x_3)/(\text{plot area}) \\ &= \text{sum of squared diameters per acre} \end{aligned}$$

If instead of a plot we have a relascope point sample, the variables s_1, s_2 and s_3 are similarly defined; only the way they are calculated is different. For more details on this, the reader is referred to Cunia (1986a). If, in addition, the plots are selected by simple random sampling and the parameter of interest is the average biomass per acre μ , the variables z_1, z_2 and z_3 are defined as the sample averages $\bar{s}_1, \bar{s}_2$ and $\bar{s}_3$ respectively. For this case, the covariance matrix $[S_{zz}]$ of $[z]$ is easy to calculate; its element ij is simply the sample covariance of $\bar{s}_i$ and $\bar{s}_j$, that is

$$\begin{aligned} S_{z_i z_j} &= S_{s_i s_j}/n \\ &= \Sigma (s_i - \bar{s}_i)(s_j - \bar{s}_j)/n(n-1) \end{aligned}$$

where n is the number of sample plots and Σ is taken over the n sample plots.

It is the objective of the present paper to extend the application of this approach to biomass estimates calculated from Continuous Forest Inventory (CFI) data of two successive measurements. The Sampling with Partial Replacement (SPR) is not being used; the case of CFI with SPR is considered in a companion paper by Cunia (1986b). We shall consider estimates of the current average biomass

⁽¹⁾ Paper based on a set of lecture notes "On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots" Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

per unit area (on the first and second measurement occasion), estimate of the average biomass per tree and estimates of growth components per unit area such as mortality, ingrowth, growth on survivor trees and net changes in biomass between the two measurements. As the difference between working with sample plots or Bitterlich sample points rests with the way the variables $s_1, s_2, \dots, s_m$ are calculated, we shall implicitly assume that whatever we say for sample plots apply equally well to sample points.

The CFI sample units are generally selected by systematic sampling but we shall assume that the bias due to the application of the simple random sampling formulae is negligibly small. It can be shown that this bias affects only the variance component associated with the sample plots. We shall also assume that the biomass tables are constructed from linear regression functions $\hat{y} = [b]'[x]$ for which the estimates of the covariance matrices $[S_{bb}]$ are given. For ways to calculate $[b]$ and $[S_{bb}]$ the reader is referred to Cunia (1986c,d) among others. The non-statistical error of using, in a given forest area, biomass tables constructed for another area is not being considered here.

We shall specifically define the vectors $[z]$ corresponding to each CFI estimator and show how to estimate their covariance matrices. Once $[z]$ and $[S_{zz}]$ are defined, the estimators w and S_{ww} can be calculated by the formulae given above. Finally, we shall illustrate the application of these formulae to a simple numerical example.

In what follows we shall use the notation

$$\begin{aligned}\hat{y}_{hk} &= b_1 x_{hk1} + b_2 x_{hk2} + \dots + b_m x_{hkm} \\ &= [b]'[x_{hk}] = \text{regression estimate of the} \\ &\quad \text{true but unknown biomass } y_{hk} \text{ of the} \\ &\quad \text{k-th tree of the h-th plot}\end{aligned}$$

and, for the case of sample plots of fixed area a ,

$$\begin{aligned}\hat{v}_h &= (\sum \hat{y}_{hk})/a = \sum (\hat{y}_{hk}/a) \\ &= \text{regression estimate of the true but un-} \\ &\quad \text{known biomass per acre } v_h \text{ of plot h,}\end{aligned}$$

where $\sum$ is taken over the trees of plot h .

When the sample unit is a Bitterlich sample point, the definition of $\hat{v}_h$ becomes

$$\begin{aligned}\hat{v}_h &= \sum (\hat{y}_{hk}/a_{hk}) \\ &= (b_1 s_{h1} + b_2 s_{h2} + \dots + b_m s_{hm})\end{aligned}$$

where a_{hk} is the plot area corresponding to the hk -th tree, and

$$s_{hi} = \sum (x_{hki}/a_{hk}) \text{ for } i = 1, 2, \dots, m$$

is the "sum of x_i values per unit area" at the sample point h .

For convenience, we may prefer using sometimes the more general notation

$$\hat{v}_h = \sum (\hat{y}_{hk}/a_{hk})$$

for both plots of fixed area (where $a_{hk} = a$ for all trees) and Bitterlich sample points. For more details on these variables the reader is sent to Cunia (1986a).

Because we shall be dealing with variables s measured on two occasions, it may be convenient to use a superscript 1 for the first and 2 for the second measurement time. For example, s_2^1 will denote variable s_2 measured on the first occasion, while s_1^2 will denote variable s_1 measured on the second occasion, and not the square of s_1 . We realize that, sometimes, this may lead to confusion. However, we shall remind the reader, whenever we feel it necessary, whether 2 is a superscript or exponent. Of course, when the superscript is not necessary, we shall simplify the notation by dropping it.

Biomass Estimators

We shall assume that (i) the CFI sample consists of n plots selected by simple random sampling, (ii) the biomass table is constructed from a linear regression function for which $[b]$ and $[S_{bb}]$ are given, and (iii) the data of the sample plots, the variables s , are statistically independent of the biomass tables.

Estimator of the Average Biomass per Unit Area μ_1 at the First Measurement Time

As the discussion is limited to the first measurement data, superscript 1 will be dropped from the notation of s_h but used only in the definition of $[z]$ and its covariance matrix $[S_{zz}]$. For notational convenience we shall drop the subscript h of the variables s_{hi} and let $\sum$ stand for summation over the values of the plot $h = 1, 2, \dots, n$.

The common procedure is to define the estimator of μ_1 as

$$\begin{aligned}w_1 &= \sum \hat{v}_h / n = (b_1 \sum s_1 + b_2 \sum s_2 + \dots + b_m \sum s_m) / n \\ &= b_1 \bar{s}_1 + b_2 \bar{s}_2 + \dots + b_m \bar{s}_m = [b]'[\bar{s}]\end{aligned}$$

If we define now

$$z_i^1 = \bar{s}_i = \sum s_i / n \text{ for } i = 1, 2, \dots, m$$

we can define $[z^1] = [\bar{s}]$ and, thus,

$$w_1 = [b]'[z^1] = \text{estimator of } \mu_1$$

Because $[b]$ and $[z^1]$ are statistically independent the variance of w_1 can be estimated by the approximate formula

$$S_{w_1 w_1} = [b]'[S_{zz}^{11}][b] + [z^1]'[S_{bb}][z^1]$$

where

$$[S_{zz}^{11}] = \text{estimator of the covariance matrix of } [z^1].$$

It remains now to calculate $[S_{zz}^{11}]$. For this we use the fact that, for $i, j = 1, 2, \dots, m$,

$$(1) \quad S_{s_i s_j} = \sum (s_i - \bar{s}_i)(s_j - \bar{s}_j) / (n-1) \\ = \text{estimator of the covariance of } s_i \text{ and } s_j$$

and

$$(2) \quad S_{\bar{s}_i \bar{s}_j} = S_{s_i s_j} / n = S_{z_i z_j}$$

If

$[S_{ss}^{11}]$ = estimator of the covariance matrix of $s_1, s_2, \dots, s_m$ as calculated from the first measurement values,

we can verify immediately that

$$[S_{zz}^{11}] = [S_{ss}^{11}] / n$$

Estimator of the Average Biomass per Unit Area μ_2 at the Second Measurement Time

The procedure is the same as in the subsection above with the second, rather than the first measurement values $s_1, s_2, \dots, s_m$ being used

If

$[z^2] = [\bar{s}]$ = vector of the average values of $s_1, s_2, \dots, s_m$ as measured on the second occasion,

and

$[S_{zz}^{22}] = [S_{ss}^{22}] / n$ = estimator of the covariance matrix of $[z^2]$

where

$[S_{ss}^{22}]$ = estimator of the covariance matrix of $s_1, s_2, \dots, s_m$ as calculated from the second measurement data,

we can verify immediately that

$$w_2 = [b]'[z^2] = \text{estimator of } \mu_2$$

and

$$S_{w_2 w_2} = [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2] \\ = \text{estimator of the variance of } w_2$$

Note that in the formulae above, 2 is a superscript not an exponent.

Estimator of the Net Change $\mu_g = (\mu_2 - \mu_1)$ in Biomass per Unit Area between the Two Measurements

The estimator of the net change can be simply defined as

$$w_g = w_2 - w_1 = [b]'[z^2] - [b]'[z^1] = [b]'[z^g]$$

where

$[z^g] = [z^2] - [z^1]$ = difference between the sample averages of $[s]$ of the first and second measurement

Because the i -th element of $[z^g]$, $i = 1, 2, \dots, m$, can be written as

$$\bar{s}_i^2 - \bar{s}_i^1 = \sum (s_{hi}^2 - s_{hi}^1) / n = \sum s_{hi}^g / n = \bar{s}_i^g$$

it suffices to work with the individual cluster values s_{hi}^g , define their averages $\bar{s}_i^g$ and covariances $S_{s_i s_j}^{gg}$, arranged as the vector of averages $[s^g]$ and covariance matrix $[S_{ss}^{gg}]$.

Then

$$[z^g] = [\bar{s}^g] \text{ and } [S_{zz}^{gg}] = [S_{ss}^{gg}] / n$$

and

$$w_g = [b]'[z^g] = \text{estimator of } \mu_g, \text{ and}$$

$$S_{w_g w_g} = [b]'[S_{zz}^{gg}][b] + [z^g]'[S_{bb}][z^g]$$

Estimators of Growth Components: μ = Average Biomass Growth per Unit Area on Survivor Trees, μ_m = Average Mortality Biomass per Unit Area and μ_i = Average Ingrowth Biomass per Unit Area

When μ_g was calculated, all trees measured on either occasion were used. If instead of all trees one uses only the survivor trees, that is, the trees that were measured on the first and were alive and measured also on the second occasion, one can define the estimator w_s of μ_s . More specifically, we define the new variables of plot h as

$$s_{hi}^s = \sum ((x_{hki}^2 - x_{hki}^1) / a_{hk}) \text{ for } i = 1, 2, \dots, m$$

where $\sum$ is taken only over the trees k of plot h that were alive and measured on both occasions. Note that x^2 denotes the value of x as measured on the second occasion and not the square of x . For the case of the relascope sample points, the "per unit area" converting factor a_{hk} is that of the first measurement. We further define

$$[z^s] = [\bar{s}^s] = \text{vector of sample averages } \bar{s}_i^s \text{ of } s_{hi}^s$$

$$[S_{ss}^{ss}] = \text{sample covariance matrix of the new plot variables } s_{hi}^s$$

and

$$[S_{zz}^{ss}] = [S_{ss}^{ss}] / n = \text{estimator of the covariance matrix of } [z^s].$$

This yields immediately

$w_s = [b]'[z^s] = \text{estimator of } \mu_s$

and

$$S_{w_s w_s} = [b]'[S_{zz}^{ss}][b] + [z_s^s]'[S_{bb}][z_s^s]$$

If in w_1 , the estimator of the average biomass per unit area of the first measurement we use only the mortality trees (the trees that were measured on the first occasion but not measured on the second because they were dead, not harvested) we obtain the estimator w_m of μ_m . More specifically, the new plot variables are

$$s_{hi}^m = \Sigma(x_{hki}^1/a_{hk}) \text{ for } i = 1, 2, \dots, m$$

where Σ is taken only over the mortality trees k of plot h. Using these new plot variables we further define

$$[z^m] = [\bar{s}^m] = \text{vector of sample averages } \bar{s}_i^m$$

$$[S_{ss}^{mm}] = \text{sample covariance matrix of } s_{hi}^m$$

$$[S_{zz}^{mm}] = [S_{ss}^{mm}]/n = \text{estimator of the covariance matrix of } [z^m]$$

$$w_m = [b]'[z^m] = \text{estimator of } \mu_m, \text{ and}$$

$$S_{w_m w_m} = [b]'[S_{zz}^{mm}][b] + [z^m]'[S_{bb}][z^m] \\ = \text{estimator of the variance of } w_m$$

Similarly, if in w_2 , the estimator of the average biomass per unit area at the second occasion, we use only the ingrowth trees (the trees that were too small to be measured on the first but became of merchantable size on the second occasion) we obtain an estimator w_i of μ_i . More specifically, we define the new plot variables

$s_{hi}^i = \Sigma(x_{hki}^2/a_{hk})$ for subscript $i = 1, 2, \dots, m$ (superscript i means ingrowth) where Σ is now taken only over the ingrowth trees k of plot h. For the case of relascope sample points, the per "unit area" factor a_{hk} is that of the second measurement time. Using the new plot variables we define the usual statistics

$$[z^i] = [\bar{s}^i] = \text{vector of sample averages } \bar{s}_i^i$$

$$[S_{ss}^{ii}] = \text{sample covariance matrix of } s_{hi}^i$$

$$[S_{zz}^{ii}] = [S_{ss}^{ii}]/n = \text{estimator of the covariance matrix of } [z^i]$$

$$w_i = [b]'[z^i] = \text{estimator of } \mu_i, \text{ and}$$

$$S_{w_i w_i} = [b]'[S_{zz}^{ii}][b] + [z^i]'[S_{bb}][z^i] \\ = \text{estimator of the variance of } w_i$$

We hope that use of superscripts s, m and i to denote survivor, mortality and ingrowth trees will not be confused with the notation s for plot variables, m for the number of variables x or s , and i for subscript of variables x_i or s_i .

Note that the estimators w_s, w_m and w_i are,

most probably biased. This is because the biomass regression function is usually defined for the entire set of all population trees. When applied to a class of trees, it is not uncommon for the biomass of the trees in this class to fall, on the average, below or above the overall regression function. For example, it is reasonable to expect the survivor trees to be the tallest and most vigorous trees of highest growth (the dominant or codominant trees) and the mortality trees to be the shortest, least vigorous trees (intermediate or suppressed). Thus, it is reasonable to assume that w_s is an underestimate and w_m an overestimate. But there is also a source of underestimation of w_m ; the growth on mortality trees between their measurement on the first occasion and the time of their death is being ignored.

It can be shown that, in the absence of harvesting, we must have (for plot, not point sampling)

$$w_2 = w_1 + w_s + w_i - w_m \text{ and } w_g = w_s + w_i - w_m$$

These relationships can be used as a check on the calculations performed.

Estimator of the Average Biomass per Tree μ_t at a Given Occasion

Because the estimation of the average biomass per tree follows the same procedure for either measurement, we shall consider the general case with no need of superscript 1 or 2.

While the plots (clusters of trees) are selected by simple random sampling, the selection of the trees themselves is made by cluster random sampling. To estimate μ_t we must therefore use the cluster sampling formulae as found in standard texts on sampling techniques as, for example, that of Cochran (1977). Using our notation, these formulae can be written as follows

$$w_t = \hat{\Sigma v}_h / \Sigma t_h = \text{estimator of } \mu_t, \text{ and}$$

$$S_{w_t w_t} = n(S_{vv} - 2w_t S_{vt} + w_t^2 S_{tt}) / (\Sigma t_h)^2 \\ = n(\Sigma v_h^2 - 2w_t \Sigma v_h t_h + w_t^2 \Sigma t_h^2) / (n-1)(\Sigma t_h)^2 \\ = \text{estimator of the variance of } w_t$$

where Σ is taken over all plots $h = 1, 2, \dots, n$ and t_h is used to denote the number of trees per unit area of plot h . Do not confuse subscript t of w_t that stands for "biomass per tree" (and may also be used as superscript) and variable t that denotes the number of trees per unit area (which may also be used as a subscript).

In this formula we have ignored the effect of the finite population correction factor; the population size is ordinarily much larger than the sample size. We have also ignored the effect of the error of the biomass tables. To take this error into account we must express first the plot biomass per unit area as

$$\hat{v}_h = b_1 s_{h1} + b_2 s_{h2} + \dots + b_m s_{hm} = [b]'[s_h]$$

Algebraic manipulation of the formula of w_t yields the formula

$$w_t = b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]'[z]^t$$

where

$z_i = \Sigma s_{hi} / \Sigma t_h$ = cluster sampling estimator of the expected value of the tree variable x_i .

The variance of z_i is estimated by the formula

$$S_{z_i z_i} = n(S_{s_i s_i} - 2z_i S_{s_i t} + z_i^2 S_{tt}) / (\Sigma t_h)^2$$

$$= n(\Sigma s_{hi}^2 - 2z_i \Sigma s_{hi} t_h + z_i^2 \Sigma t_h^2) / (n-1)(\Sigma t_h)^2$$

and the covariance of z_i and z_j by the formula

$$S_{z_i z_j} = n(S_{s_i s_j} - z_i S_{s_j t} - z_j S_{s_i t} + z_i z_j S_{tt}) / (\Sigma t_h)^2$$

$$= n(\Sigma s_{hi} s_{hj} - z_i \Sigma s_{hj} t_h - z_j \Sigma s_{hi} t_h + z_i z_j \Sigma t_h^2) / (n-1)(\Sigma t_h)^2$$

Using the formulae above, we calculate

$$[S_{zz}^{tt}] = \text{estimator of the covariance matrix of } [z]^t$$

This yields immediately

$$w_t = [b]'[z]^t = \text{estimator of the average biomass per tree}$$

and

$$S_{w_t w_t} = [b]'[S_{zz}^{tt}][b] + [z]^t[S_{bb}][z]^t$$

= estimator of the variance of w_t

It may be interesting to see what happens when $z_i \equiv 1$. This may be the case when the sample unit is a plot of fixed area "a"

and

$$s_{h1} = \Sigma (1/a) = \text{number of trees per unit area} = t_h$$

where Σ is taken over all trees of a given plot h.

Then,

$$z_1 = \Sigma s_{h1} / \Sigma t_h \equiv 1$$

and it can be easily shown that, for all $i = 1, 2, \dots, m$

$$S_{z_1 z_i} = 0$$

Illustrative Numerical Example

To see how to apply the formulae of the previous section, we shall work with data from 118 permanent, one-tenth acre sample plots selected at random from the Maniwaki region of Québec-

Canada. These plots were measured on two occasions in 1960 and 1964. To estimate the biomass of sample trees and plots, we shall use a tree biomass regression function of the parabolic form

$$y = b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x]$$

where $x_1 = 1$, x_2 = (tree diameter) and x_3 = (squared tree diameter).² This regression function was calculated from a random sample of 353 trees by Cunia (1986c) and the following statistics are given.

$$[b] = \begin{bmatrix} 5.1818118 \\ -25.653078 \\ 12.988357 \end{bmatrix} \quad \text{and}$$

$$[S_{bb}] = \begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

The biomass is given in pounds of above ground green weight and the tree diameter is measured in inches.

The plot variables corresponding to this regression function are

s_1 = number of trees/acre

s_2 = sum of tree diameters/acre, and

s_3 = sum of squared tree diameters/acre

Sometimes, a superscript 1 for the first and 2 for the second measurement may be added. It should not be confused with exponents 1 or 2.

Estimation of μ_1 , μ_2 and $\mu_g = (\mu_2 - \mu_1)$

The values of the individual plot variables s_1 , s_2 , and s_3 as measured in 1960 and 1964 on the $n = 118$ permanent plots of our sample are given in Cunia (1986e), and the values of their means, variances and covariances are shown below.

$$[s^1] = \begin{bmatrix} 236.35593 \\ 1439.3161 \\ 10898.164 \end{bmatrix} = \text{sample mean of } [s] \text{ of the first measurement}$$

$$[s^2] = \begin{bmatrix} 266.01695 \\ 1638.6805 \\ 12320.176 \end{bmatrix} = \text{sample mean of } [s] \text{ of the second measurement}$$

$$[S_{ss}^{11}] = \begin{bmatrix} 21272.932 & 110452.37 & 602351.30 \\ 110452.37 & 659011.32 & 4724541.6 \\ 602351.30 & 4724541.6 & 48604076 \end{bmatrix}$$

= sample covariance matrix of $[s]$ of the first measurement

$$[S_{ss}^{22}] = \begin{bmatrix} 24039.555 & 123557.78 & 635935.23 \\ 123557.78 & 718316.31 & 4802637.7 \\ 635935.23 & 4802637.7 & 47000205 \end{bmatrix}$$

= sample covariance matrix of $[s]$ of the second measurement

$$[S_{ss}^{12}] = \begin{bmatrix} 21987.926 & 118295.97 & 659678.73 \\ 109292.82 & 670898.34 & 4832348.2 \\ 542577.80 & 4443890.0 & 46552860 \end{bmatrix}$$

= sample covariance matrix of [s] of the first with [s] of the second measurement

Note that $[S_{ss}^{12}]$ is not symmetrical and that the covariance terms associated with s_1 of the first measurement are found in row 1, while the covariance terms associated with s_1 of the second measurement are found in column 1.

These statistics are sufficient to estimate μ_1 , μ_2 and net change μ_g and their variances. Using the formulae of the previous section, the reader can verify that

$$[z^1] = [\bar{s}^1] = \begin{bmatrix} 236.35593 \\ 1439.3161 \\ 10898.164 \end{bmatrix},$$

$$[z^2] = [\bar{s}^2] = \begin{bmatrix} 266.01695 \\ 1638.6805 \\ 12320.176 \end{bmatrix}$$

$$[z^g] = [z^2] - [z^1] = \begin{bmatrix} 29.661017 \\ 199.36441 \\ 1422.0114 \end{bmatrix}$$

$$[S_{zz}^{11}] = [S_{ss}^{11}]/n = \begin{bmatrix} 180.27909 & 936.03701 & 5104.6720 \\ 936.03701 & 5584.8417 & 40038.488 \\ 5104.6720 & 40038.488 & 411898.95 \end{bmatrix}$$

$$[S_{zz}^{22}] = [S_{ss}^{22}]/n = \begin{bmatrix} 203.72504 & 1047.0999 & 5389.2817 \\ 1047.0999 & 6087.4263 & 40700.319 \\ 5389.2817 & 40700.319 & 398306.82 \end{bmatrix}$$

and

$$[S_{zz}^{gg}] = ([S_{ss}^{11}] - [S_{ss}^{12}] - [S_{ss}^{12}]' + [S_{ss}^{22}])/n$$

$$= \begin{bmatrix} 11.327426 & 54.418338 & 305.33909 \\ 54.418338 & 301.10981 & 2126.6194 \\ 305.33909 & 2126.6194 & 21174.239 \end{bmatrix}$$

This yields immediately the estimates

$$w_1 = [b]'[z^1] = 105851.11 \text{ pounds}$$

= estimate of the mean biomass per acre at the first, 1960 measurement

$$w_2 = [b]'[z^2] = 119360.09 \text{ pounds}$$

= estimate of the mean biomass per acre at the second, 1964 measurement

$$w_g = [b]'[z^g] = 13508.98 \text{ pounds} = w_2 - w_1$$

= estimate of the mean change in biomass per acre between the two measurements

$$S_{w_1 w_1} = [b]'[S_{zz}^{11}][b] + [z^1]'[S_{bb}][z^1]$$

$$= 46923716 + 10710584 = 57634300$$

= estimate of the variance of w_1

$$S_{w_2 w_2} = [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2]$$

$$= 44529880 + 11393835 = 55923716$$

= estimate of the variance of w_2

and

$$S_{w_g w_g} = [b]'[S_{zz}^{gg}][b] + [z^g]'[S_{bb}][z^g]$$

$$= 2379986 + 50666 = 2430652$$

= estimate of the variance of w_g

The variance of each estimator above has two main components, the first associated with the variation between sample plots, the second associated with the biomass table.

We shall now show that the common procedure consisting of the calculation of (i) the individual plot biomass per acre $\hat{v}_h$, (ii) average biomass per acre $\bar{v}$ and (iii) the variance of $\bar{v}$ as the variance of individual plot values $\hat{v}_h$ divided by n , yields the first variance component. Using the individual plot biomass per acre as given by Cunia (1986e), we find

$$\Sigma \hat{v}_h = 12490431 \text{ for the first measurement}$$

$$= 14084491 \text{ for the second measurement}$$

$$\Sigma (\hat{v}_h)^2 = 196995482800 \text{ for the first measurement}$$

$$= 2295905587000 \text{ for the second measurement}$$

and the sum of the cross products between the first and the second measurement values $\hat{v}_h$, as equal to 2105734088000.

The reader can verify then that

$$S_{vv}^{11} = 5536998459, S_{vv}^{22} = 5254525850, S_{vv}^{12} = 5255342828$$

$$w_1 = \bar{v} \text{ of the first measurement}$$

$$= 12490431/118 = 105851.11$$

$$w_2 = \bar{v} \text{ of the second measurement}$$

$$= 14084491/118 = 119360.09$$

$$w_g = w_2 - w_1 = 13508.98$$

$$S_{w_1 w_1} = S_{vv}^{11}/n = 5536998459/118 = 46923716$$

$$S_{w_2 w_2} = S_{vv}^{22}/n = 5254525850/118 = 44529880$$

and

$$S_{w_g w_g} = (S_{vv}^{11} - 2S_{vv}^{12} + S_{vv}^{22})/118 = 2379986$$

Let us now see the percent of the total

error represented by the error of the biomass regression function. When estimating μ_1 and μ_2 the percentages are 18.58 and 20.37. The percentage becomes negligibly small, about 2 percent when estimating the net change μ_g . It appears that the effect of the error of the biomass regression may be ignored when the net change is estimated but the effect is sizable when the current averages are estimated. In this last case, it is also important to realize that we have a large sample of trees (from which the biomass regression function is estimated) and a small sample of plots. For illustration purposes let us consider a much more realistic sample of 88 trees (one fourth of our sample of 353 trees) and a sample of 472 plots (four times as many as our sample of 118 plots). Then, the variance of w_1 is expected to be about

$$(46923716/4) + (4) (10710584)$$

$$= 11730929 + 42842336 = 54573265$$

and the percent of the total error due to the biomass regression is now about 78.50, an extremely large value.

Estimation of Growth Components μ_s , μ_m and μ_i

To estimate the average biomass growth on survivors and the biomass of the mortality and ingrowth trees, one should go to the original tree measurements and calculate the new plot values of $[s^s]$, $[s^m]$ and $[s^i]$. These new values are given by Cunia (1986e) and not repeated here. Using them, we calculate the following vectors of averages and the covariance matrices

$$[s^s] = \begin{bmatrix} 0 \\ 107.75 \\ 1346.0135 \end{bmatrix}, \quad [s^m] = \begin{bmatrix} 7.2033898 \\ 49.911017 \\ 470.95899 \end{bmatrix}$$

$$[s^i] = \begin{bmatrix} 36.864407 \\ 141.52542 \\ 546.95690 \end{bmatrix}$$

$$[S_{ss}^{ss}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3371.6046 & 36266.937 \\ 0 & 36266.937 & 437765.17 \end{bmatrix}$$

$$[S_{ss}^{mm}] = \begin{bmatrix} 148.52238 & 1066.5183 & 11480.546 \\ 1066.5183 & 9533.2480 & 123761.82 \\ 11480.546 & 123761.82 & 1840636.3 \end{bmatrix}$$

$$[S_{ss}^{ii}] = \begin{bmatrix} 1120.8533 & 4304.0120 & 16653.079 \\ 4304.0120 & 16595.353 & 64517.716 \\ 16653.079 & 64517.716 & 252214.66 \end{bmatrix}$$

Using the relationships $[z] = [S]$ and $[S_{zz}] = [S_{ss}]/n$, we find that

$$w_s = [b]'[z^s] = 14718.385 \text{ pounds}$$

= estimate of the biomass growth per acre on survivor trees

$$w_m = [b]'[z^m] = 4873.939 \text{ pounds}$$

= estimate of the biomass per acre loss

due to mortality

$$w_i = [b]'[z^i] = 3664.5330 \text{ pounds}$$

= estimate of the biomass per acre growth due to ingrowth trees

$$S_{w_s w_s} = [b]'[S_{zz}^{ss}][b] + [z^s]'[S_{bb}][z^s]$$

$$= 439839 + 608965 = 1048804$$

= estimate of the variance of w_s

$$S_{w_m w_m} = [b]'[S_{zz}^{mm}][b] + [z^m]'[S_{bb}][z^m]$$

$$= 1996418 + 24567 = 2020984$$

= estimate of the variance of w_m

and

$$S_{w_i w_i} = [b]'[S_{zz}^{ii}][b] + [z^i]'[S_{bb}][z^i]$$

$$= 98330 + 767722 = 866052$$

= estimate of the variance of w_i

It can be verified that

$$w_2 = w_1 + w_s + w_i - w_m = 105851 + 14718 + 3665 - 4874 = 119360$$

and

$$w_g = w_s + w_i - w_m = 14718 + 3665 - 4874 = 13509$$

The percent of the total error of the estimate of mortality, due to the error of the biomass regression function is negligible, about 1.22; this may be due to the fact that the variation of mortality from plot to plot is very large. On the other hand the percent is very high for survivor growth and ingrowth about 58.06 and 88.65 respectively. In the first case the pairs of values of the survivor trees are very highly correlated and thus, their differences have a small variance, and in the second case, the ingrowth trees are small in size, they are all estimated from the extreme side of the biomass regression function where the error is extremely large.

Estimation of the Average Biomass per Tree μ_t

We shall start with the calculation of the average biomass per tree for the first measurement. For notational convenience, the superscript 1, and later 2, will not be shown, and subscript h will be dropped. The set of original data has been summarized as the following basic statistics. Recall that $v_h = [b]'[s_h]$ is the biomass per acre of plot h, and $t_h = s_{h1}$ is the number of trees per acre of plot h.

$$\Sigma v = 12490431, \Sigma t = \Sigma s_1 = 27890, \Sigma s_2 = 169839.3,$$

$$\Sigma s_3 = 1285983.399, S_{vv} = 5536998459,$$

$$S_{vt} = 5100342.7, S_{tt} = 21272.932 (=S_{s_1 s_1})$$

The covariance matrix of [s] has already been given as $[S_{ss}^{11}]$ when μ_1 was estimated and

$$[S_{st}] = \begin{bmatrix} s_{1t} \\ s_{2t} \\ s_{3t} \end{bmatrix} = \begin{bmatrix} s_{s_1 s_1} \\ s_{s_1 s_2} \\ s_{s_1 s_3} \end{bmatrix} = \begin{bmatrix} 21272.932 \\ 110452.37 \\ 602351.30 \end{bmatrix}$$

= sample covariance vector of [s] and t.

Using matrix notation, the reader can verify that

$$[z^t] = \begin{bmatrix} \Sigma s_1 \\ \Sigma s_2 \\ \Sigma s_3 \end{bmatrix} / \Sigma t = \begin{bmatrix} 1 \\ 6.0896128 \\ 46.109122 \end{bmatrix}$$

and

$$[S_{zz}^{tt}] = n([S_{ss}^{11}] - [z^t][S_{st}]' - [S_{st}][z^t]' + [z^t][z^t]') / (\Sigma t)^2$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 0 & .015573400 & .29380459 \\ 0 & .29380459 & 5.8076198 \end{bmatrix}$$

For example,

$$S_{z_2 z_3}^{tt} = (118)(4724541.6) - (6.0896128)(602351.30) - (46.109122)(110452.37) + (6.0896128)(46.109122)(21272.932) / (27890)^2 = .29380459$$

This yields immediately

$$w_t = [b]'[z^t] = 447.84622 \text{ pounds}$$

= estimate of the average biomass per tree at the first measurement

and

$$S_{w_t w_t} = [b]'[S_{zz}^{tt}][b] + [z^t]'[S_{bb}][z^t] = 794.19269 + 191.72561 = 985.91831$$

= estimate of the variance of w_t

It can be easily verified that the first component of the variance of w_t is that given by the common procedure

$$S_{w_t w_t} = n(s_{vv} - 2w_t s_{vt} + w_t^2 s_{tt}) / (\Sigma t)^2 = (118)(5536998459 - (2)(447.84622)(5100342.7) + (447.84622)^2(21272.932)) / (27890)^2 = 794.19269$$

since

$$w_t = \Sigma v / \Sigma t = 12490431 / 27890 = 447.84622$$

Similarly, to estimate the average biomass per tree at the second measurement time we start with the basic statistics

$$\Sigma t = \Sigma s_1 = 31390, \Sigma s_2 = 193364.3,$$

$$\Sigma s_3 = 1453780.747, S_{tt} = 24039.555$$

$$[S_{ss}] = [S_{ss}^{22}] \text{ as given before and}$$

$$[S_{st}] = \begin{bmatrix} 24039.555 \\ 123557.78 \\ 635935.23 \end{bmatrix}$$

Then the reader can verify that

$$[z^t] = \begin{bmatrix} 1 \\ 6.1600605 \\ 46.313499 \end{bmatrix},$$

$$[S_{zz}^{tt}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & .012967567 & .24205016 \\ 0 & .24205016 & 4.7494072 \end{bmatrix}$$

$$w_t = [b]'[z^t] = 448.69355$$

= estimate of the average biomass per tree at the second measurement

and

$$S_{w_t w_t} = [b]'[S_{zz}^{tt}][b] + [z^t]'[S_{bb}][z^t] = 648.44837 + 161.00944 = 809.45780$$

= estimate of the variance of w_t

As a simple check of some of our calculations, we find

$$w_t(\Sigma t) / n = 105851.11 = w_1 = \text{for the first measurement} \\ = 119360.09 = w_2 = \text{for the second measurement}$$

Extension to More than One Species

The formulae above can be easily extended to more than one species. We start with the giant size vector [B] of the regression coefficients vectors $[b^1], [b^2], \dots, [b^p]$ of species 1, 2, ..., p respectively, that is,

$$[B]' = [[b^1]' [b^2]' \dots [b^p]']$$

and the covariance matrices $[S_{bb}^{11}], [S_{bb}^{22}], \dots, [S_{bb}^{pp}]$

arranged in the giant size covariance matrix

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [0] & \dots & [0] \\ [0] & [S_{bb}^{22}] & \dots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \dots & [S_{bb}^{pp}] \end{bmatrix}$$

It is assumed that $[b^i]$ is statistically independent of $[b^j]$, for $i \neq j = 1, 2, \dots, p$. If not, one should calculate all the covariance matrices $[S_{ij}^{ij}]$ of $[b^i]$ and $[b^j]$ and complete the matrix $[S_{bb}^{ij}]$ above. One way in which $[S_{bb}^{ij}] \neq 0$ may arise

is shown by Cunia (1986d).

For each species we calculate the plot vector $[s]$, say $[s^1]$, $[s^2]$, ..., $[s^p]$ of species 1, 2, ..., p respectively. After constructing the giant size vector

$$[s_h]' = [[s_h^1]' [s_h^2]' \dots [s_h^p]']$$

for each plot h, we calculate, by the usual formulae

$$[Z] = [\bar{S}] = \text{vector of the averages of } [s_h]$$

and

$$[S_{ZZ}] = (\text{covariance matrix of } [s_h])/n$$

Then,

$$w_1 = [B]'[Z] = \text{estimator of } \mu_1 \text{ (of all species)}$$

and

$$\begin{aligned} s_{w_1 w_1} &= [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z] \\ &= \text{estimator of the variance of } w_1 \end{aligned}$$

The same approach can be applied to estimate the other parameters of interest, such as, for example, μ_2 , μ_g or μ_t .

Concluding Remarks

The common procedure for the calculation of the error of volume estimates from Continuous Forest Inventory data ignores the effect of the error of the tree biomass tables; it is implicitly assumed that this error is negligibly small. And this may indeed be true whenever (i) the number of CFI plots or points is relatively small, (ii) the sample of trees from which the biomass tables was constructed is relatively large and (iii) we are concerned with estimators of current volumes per unit area or per tree. Then, the error of the sample plots is much larger than the error of the biomass tables.

We have reconsidered the approach proposed by Cunia (1965, 1986a) to add the error of biomass tables to the error of sample plots, when the total error of biomass estimates is calculated. We have extended the applicability of this approach to CFI systems for estimates of (i) average biomass per unit area, (ii) average biomass per tree and (iii) average biomass per unit areas for growth components such as net change between measurements, mortality, ingrowth or survivor growth. We have also shown how to extend the applicability of the formulae from plots of fixed area to Bitterlich sample points.

We have then illustrated how to apply these formulae to an actual case of 118 permanent sample plots and a biomass regression function calculated from 353 sample trees. We have seen that, with this combination of sample sizes, the percent of the total error associated with the biomass tables is only about 20 percent, when the estimates of average biomass per acre at the

first and second measurement are calculated. The percent remains the same for the average biomass per tree. We felt, however, that these sample sizes are not representative of what is actually occurring in real life. Working with four times as many plots (472) and four times as few sample trees (88) we have shown that the expected part of the error of the biomass tables may rise to about 80 percent. This size of the error is indeed very high and can hardly be ignored.

The effect of the error of biomass regression function may be quite different when we consider estimates of growth components. It may be negligibly small for estimates of net change and mortality and considerably large for estimates of survivor growth and ingrowth. Sometimes this can be explained. The ingrowth is calculated from small trees, whose biomass regression estimates have a large error; they are all calculated from the extreme of the regression function where the error is at its maximum. In the case of survivor growth, the error due to sample plots is small (relative to that of biomass tables) because the differences between the two measurements of the same trees have a very small variance; the correlation between the two measurements is extremely high. We are at a loss, however, to explain why the part of the error due to biomass regression function is small for net change and mortality; it is possible that the variation of the net change and mortality from plot to plot overshadows completely in size the error of biomass tables.

We have completely ignored the possible bias of the biomass tables calculated for one, and applied to another forest area. It is assumed that the two forest areas are sufficiently similar for the bias to be small. However, this may not be true. Because the bias is non-statistical in nature and can hardly be estimated, it is generally ignored. By properly designing (i) the sampling method for selecting sample trees for biomass regression function and (ii) the statistical procedure by which the regression function is calculated, one may be able to quantify the error component due to the difference between forest areas or stands, and, thus, take it into account when calculating the total error of the biomass estimates. More research is, however, needed in this area.

We have finally shown that the extension of the methodology from one to several species is straightforward. It requires the construction of (i) giant size vectors $[B]$ containing the individual species vectors $[b]$ and their covariance matrix $[S_{BB}]$ and (ii) giant size vectors of sample plot values $[s]$ corresponding to vector $[B]$ and their averages $[\bar{S}] = [Z]$ and covariance matrices $[S_{ZZ}] = [S_{ss}]/n$. Then, the formulae for the estimator w of μ and its variance are the same.

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

Literature Cited

Cochran, W.G. Sampling Techniques, 3rd Ed. John Wiles and Sons, New York, NY; 1977.

Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the workshop on "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; 1986a.

Cunia, T. On the error of forest inventory estimates: Continuous Forest Inventory with SPR. In: Proceedings of the workshop on "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; 1986b.

Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "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; 1986c.

Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the workshop on "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; 1986d.

Cunia, T. On the error of biomass estimates in forest inventories: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (86-001), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986e.

ON THE ERROR OF FOREST INVENTORY ESTIMATES:

CONTINUOUS FOREST INVENTORY WITH SPR. (1)

Tiberius Cunia

Professor of Statistics and Operations Research
SUNY College of Environmental Science and Forestry,
Syracuse, NY, 13210

In a previous paper, Cunia has shown how to take into account the error due to biomass tables when the error of forest biomass estimates is calculated in Continuous Forest Inventory (CFI). This approach is extended to the case of CFI systems using Sampling with Partial Replacement (SPR) methodology. The formulae to calculate the error of current biomass estimates and the estimates of growth components are given for the case of SPR on two successive measurements when (i) biomass tables are constructed from linear regression functions for which an estimate of the covariance matrix of the regression coefficients is given and (ii) the CFI plots or relascope sample points are selected by simple random sampling independently of the given biomass regression functions.

Introduction

Although it is known that the biomass regressions applied to forest inventory are not without error, it is a common procedure to ignore this error when inferences are made about the average biomass per unit area. Cunia (1986a) has proposed one approach to combine the error from inventory sample plots with the error from biomass regressions when (i) the regressions are linear and the vector $[\beta]$ of m coefficients is estimated by the vector $[b]$ calculated by least squares techniques, (ii) there are m statistics $z_1, z_2, \dots, z_m$ calculated from the sample plots data, so that the vector $[z]$ of their expected values multiplied by the vector $[\beta]$ is equal to μ , the parameter one wishes to estimate and (iii) the vectors $[b]$ and $[z]$ are statistically independent. He gave the formulae for the estimators

$$w = [b]'[z] \text{ of the parameter of interest } \mu,$$

and

$$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$$

of the variance of w , where $[S_{zz}]$ and $[S_{bb}]$ are the known estimators of the covariance matrices of $[z]$ and $[b]$ respectively. Note that the left hand side of the expression of the variance

(1) Paper based on a set of lecture notes "On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots". Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

of w may be viewed as an error component due to sample plots, while the right hand side may be viewed as the error component due to biomass regressions.

In his paper, Cunia (1986a) considers the case where the sample plots are selected by simple random sampling and the parameter of interest μ is the mean biomass per unit area. This procedure was then extended, see Cunia (1986a), to the case where the sample plots were all permanent and part of a Continuous Forest Inventory (CFI) system and the parameters of interest are the mean biomass per acre at the first and second occasion (μ_1 and μ_2), the net change in the mean biomass per acre between the first and second occasion ($\mu_g = \mu_2 - \mu_1$), the mean biomass per tree (μ_t) and the mean biomass growth per acre by growth component (μ_m for mortality, μ_i for in-growth and μ_s for growth on survivor trees).

It is known that CFI systems can be made more efficient when, at each measurement time, Sampling with Partial Replacement (SPR) methodology is being used and a part of the old sample plots is being replaced by new plots. A relationship between past and present measurement values is established from the data of the plots that are being remeasured, and this relationship is being used to update the values of the old plots that are not being remeasured and backdate the values of the new plots that have not been measured in the past.

It is the objective of the present paper to derive the formulae for the estimation of (i) current values and growth for CFI systems using SPR, and (ii) the error of these estimates that includes the error of biomass regressions in addition to the error of the CFI plots or points. For the general approach of how to combine the error of biomass regressions with that of the sample plots, the reader is sent back to Cunia (1986a,b). For the general methodology of SPR as applied to Forest Inventory the reader is sent to papers by Ware and Cunia (1962), Cunia (1965), Cunia and Chevrou (1969) and Newton, Cunia and Bickford (1974) among others.

We shall use the terminology and notation introduced by Cunia (1986a,b), in particular that of the cluster (plot) variables $s_1, s_2, \dots$, defined on a "per unit area, acre or hectare" basis. For convenience, these variables will be defined in a later section. Because these variables can be defined on both sample plots of fixed area or relascope sample points, it is immaterial whether CFI sample units are plots or points; for convenience we shall use plots to denote them both.

Only CFI systems with SPR on two occasions will be considered here. To facilitate the derivation of the methodology that combines the error from sample plots and volume regressions, we shall present the SPR formulae in a more streamlined matrix form which may differ from that used in the above-mentioned SPR papers.

The Basic Methodology of SPR on Two Occasions

We shall start with a summary description of the basic methods of calculating the SPR estimators of μ_1 = average biomass per unit area at the first measurement time, μ_2 = average biomass per unit area at the second measurement time, and $\mu_g = \mu_2 - \mu_1$ = average change in biomass from the first to the second measurement time. We shall assume that there are $(u + m + n)$ sample plots selected by simple random sampling, where u (for unmatched) plots were measured on the first but not on the second occasion, m (for matched) plots were measured on both occasions, and n (for new) plots were measured on the second but not on the first occasion.

To describe the methodology, we shall need additional notation. The biomass regression functions are usually defined in terms of dependent and independent variables y and x respectively. Because we work only with vector $[b]$ and matrix $[S_{bb}]$, we shall use x and y to denote different values. More specifically, we shall use variables $x_1, x_2, \dots, x_p$ to denote measurements made on sample plots on the first occasion, and variables $y_1, y_2, \dots, y_q$ to denote measurements made on sample plots on the second occasion. For convenience, we shall arrange these variables as vectors $[x]$ and $[y]$. Except for x_1 and y_1 defined as biomass per unit area on the first and second occasion respectively, the other x and y variables (which we shall call here auxiliary) may or may not be paired with each other.

Let the vector of the averages of $x_1, x_2, \dots, x_p$, as calculated by the usual formulae from the u unmatched plots be denoted as

$$[\bar{x}_u]' = [\bar{x}_{1u} \bar{x}_{2u} \dots \bar{x}_{pu}]$$

where $[]'$ means transpose of $[]$. Similarly, the vector of the averages of $y_1, y_2, \dots, y_q$, as calculated from the n new plots is denoted as

$$[\bar{y}_n]' = [\bar{y}_{1n} \bar{y}_{2n} \dots \bar{y}_{qn}]$$

and the vectors of averages of the m matched plots will be denoted as

$$[\bar{x}_m]' = [\bar{x}_{1m} \bar{x}_{2m} \dots \bar{x}_{pm}]$$

and

$$[\bar{y}_m]' = [\bar{y}_{1m} \bar{y}_{2m} \dots \bar{y}_{qm}]$$

The sample estimators of the variances and covariances of x and y variables are derived from the m matched plots only, by the usual formulae. Denoted as $S_{x_1x_1}, S_{x_1y_1}$, and $S_{y_1y_1}$ they are arranged

in the three covariance matrices $[S_{xx}]$ of $[x]$, $[S_{xy}]$ of $[x]$ and $[y]$ and $[S_{yy}]$ of $[y]$. Note that the order of $[S_{xy}]$ is p by q , since the covariance terms of a given variable x_i are arranged in the i -th row, while the covariance terms of a given variable y_j are arranged in the j -th column.

The SPR estimators, say $\hat{x}_1$ of μ_1 , $\hat{y}_1$ of μ_2 ,

and $\hat{g}$ of $\mu_g = (\mu_2 - \mu_1)$ are now defined as the 3 by 1 vector

$$[w] = \begin{bmatrix} \hat{x}_1 \\ \hat{y}_1 \\ \hat{g} \end{bmatrix} = [Q] + [A]'[P]$$

where

$$[Q] = \begin{bmatrix} \bar{x}_{1u} \\ \bar{y}_{1n} \\ \bar{y}_{1n} - \bar{x}_{1u} \end{bmatrix} = \text{vector of the averages of unmatched (and new) plots}$$

$$[P] = \begin{bmatrix} [\bar{x}_u] - [\bar{x}_m] \\ [\bar{y}_n] - [\bar{y}_m] \end{bmatrix} = \text{vector of the } (p + q) \text{ differences between the averages of the matched and unmatched (and new) plots}$$

$[A] = [G]^{-1}[H]$ = the $(p + q)$ by 3 matrix of the SPR coefficients

$$[G] = \begin{bmatrix} \left(\frac{u+m}{um}\right)[S_{xx}] & \left(\frac{1}{m}\right)[S_{xy}] \\ \left(\frac{1}{m}\right)[S_{xy}]' & \left(\frac{m+n}{mn}\right)[S_{yy}] \end{bmatrix}$$

= covariance matrix of $[P]$

$[G]^{-1}$ = inverse of $[G]$

and

$$[H] = \begin{bmatrix} S_{x_1x_1}/u & 0 & -S_{x_1x_1}/u \\ \vdots & \vdots & \vdots \\ S_{x_1x_p}/u & 0 & -S_{x_1x_p}/u \\ 0 & S_{y_1y_1}/n & S_{y_1y_p}/n \\ \vdots & \vdots & \vdots \\ 0 & S_{y_1y_q}/n & S_{y_1y_q}/n \end{bmatrix}$$

= covariance matrix of $[P]$ and $[Q]$

If we define, in addition

$$[K] = \begin{bmatrix} S_{x_1x_1}/u & 0 & -S_{x_1x_1}/u \\ 0 & S_{y_1y_1}/n & S_{y_1y_1}/n \\ -S_{x_1x_1}/u & S_{y_1y_1}/n & (S_{x_1x_1}/u + S_{y_1y_1}/n) \end{bmatrix}$$

then, the estimator of the covariance matrix of $[w]$ is

$$[S_{ww}] = [K] + [A]'[H] = [K] - [H]'[G]^{-1}[H]$$

Various SPR estimators are generally obtained whenever different sets of auxiliary variables are being used. All these estimators ignore the effect of the biomass regression function used in the calculation of x_1 and y_1 . To take this effect into account, we may use the approach of the next section.

The New SPR Estimators

To take into account the error of the volume tables, we need plot variables $s_1, s_2, \dots$ of the type defined by Cunia (1986 a,b). For more details the reader is referred to these papers. It suffices to say here that if the regression function of u = tree biomass on some independent variables $v_1, v_2, \dots, v_r$ is of the form

$$\hat{u} = b_1 v_1 + b_2 v_2 + \dots + b_r v_r$$

then, the plot variables are defined as

$$s_i = \Sigma(v_{ik}/a_k), \quad i = 1, 2, \dots, r$$

where Σ is taken over all trees k in a plot or point and a_k is the factor that converts the measurement v_{ik} of tree k to a "per unit area" basis. For example, if $v_1 = 1$ and $v_2 = d$ = tree diameter, and the tree is selected from a plot of fixed area of "a" acres, then

$$s_1 = \Sigma(v_1/a) = \Sigma(1/a) = \text{number of trees per acre, and}$$

$$s_2 = \Sigma(v_2/a) = \Sigma(d/a) = \text{sum of diameters per acre}$$

We start with the same three samples of u unmatched, m matched and n new plots. Each plot is measured for the variables $s_1, s_2, \dots, s_r$ on the first, on the second, or on both occasions depending on whether it is one of the u , one of the n or one of the m plots respectively. Using wherever necessary superscripts to denote the first (1) or the second (2) measurement, and subscript u, m and n to denote the sample from which the statistics are calculated, we define

(1) the vectors of sample averages

$$[\bar{s}_u]' = [\bar{s}_{1u} \quad \bar{s}_{2u} \quad \dots \quad \bar{s}_{ru}]$$

$$[\bar{s}_n]' = [\bar{s}_{1n} \quad \bar{s}_{2n} \quad \dots \quad \bar{s}_{rn}]$$

$$[\bar{s}_m^1]' = [\bar{s}_{1m}^1 \quad \bar{s}_{2m}^1 \quad \dots \quad \bar{s}_{rm}^1] \text{ and}$$

$$[\bar{s}_m^2]' = [\bar{s}_{1m}^2 \quad \bar{s}_{2m}^2 \quad \dots \quad \bar{s}_{rm}^2]$$

(2) the covariance matrices of $[s^1]$ and $[s^2]$ calculated from the data of the m matched plots alone,

$$[S_{ss}^{11}] = \text{sample covariance matrix of } [s^1], \text{ the vector of the first measurement values } s_1, s_2, \dots, s_r$$

$$[S_{ss}^{12}] = \text{sample covariance matrix of } [s^1], \text{ the vector of the first and } [s^2], \text{ the}$$

vector of the second measurement values $s_1, s_2, \dots, s_r$

and

$$[S_{ss}^{22}] = \text{sample covariance matrix of } [s^2], \text{ the vector of the second measurement values } s_1, s_2, \dots, s_r$$

Note (i) the use of superscripts 1 and 2 for the statistics of the m matched plots and the dropping of these superscripts when they are not needed, (ii) the covariance terms of $[S_{ss}^{12}]$ associated with s_i of the first measurement are all on the row i , while those associated with s_j of the second measurement are all on the column j , and (iii) $[S_{ss}^{21}] = [S_{ss}^{12}]'$.

Consider now the following vectors

$$[Q_1] = [\bar{s}_u], [Q_2] = [\bar{s}_n], [Q_g] = [Q_2] - [Q_1]$$

and

$$[P] = \begin{bmatrix} [\bar{s}_u] - [\bar{s}_m^1] \\ [\bar{s}_n] - [\bar{s}_m^2] \end{bmatrix}$$

The covariance matrices of these vectors are

$$[K_1] = [S_{ss}^{11}]/u = \text{estimator of the covariance matrix of } [Q_1]$$

$$[K_2] = [S_{ss}^{22}]/n = \text{estimator of the covariance matrix of } [Q_2]$$

$$[K_g] = [K_1] + [K_2] = \text{estimator of the covariance matrix of } [Q_g]$$

$$[G] = \begin{bmatrix} \left(\frac{u+m}{um}\right) [S_{ss}^{11}] & \left(\frac{1}{m}\right) [S_{ss}^{12}] \\ \left(\frac{1}{m}\right) [S_{ss}^{12}]' & \left(\frac{m+n}{mn}\right) [S_{ss}^{22}] \end{bmatrix}$$

= estimator of the covariance matrix of $[P]$

$$[H_1] = \begin{bmatrix} [K_1] \\ [0] \end{bmatrix} = \text{estimator of the covariance matrix of } [P] \text{ and } [Q_1]$$

$$[H_2] = \begin{bmatrix} [0] \\ [K_2] \end{bmatrix} = \text{estimator of the covariance matrix of } [P] \text{ and } [Q_2]$$

and

$$[H_g] = \begin{bmatrix} -[K_1] \\ [K_2] \end{bmatrix} = \text{estimator of the covariance matrix of } [P] \text{ and } [Q_g]$$

where

$[0]$ is the r by r matrix of zero.

We can define now the SPR estimators of the expected values of $s_1, s_2, \dots, s_r$ denoted here as the mean vectors $[\mu_z^1]$ for the first measurement, $[\mu_z^2]$ for the second measurement, and $[\mu_z^g] = [\mu_z^2] - [\mu_z^1]$ for the net change from the first to the second

measurement as the statistics

$$[z^1] = [Q_1] + [A_1]'[P]$$

$$[z^2] = [Q_2] + [A_2]'[P]$$

and

$$[z^g] = [Q_g] + [A_g]'[P]$$

where

$$[A_1] = [G]^{-1}[H_1]$$

$$[A_2] = [G]^{-1}[H_2]$$

and

$$[A_g] = [G]^{-1}[H_g] = [A_2] - [A_1]$$

The covariance matrices of $[z^1]$, $[z^2]$ and $[z^g]$ are estimated by the formulae

$$[S_{zz}^{11}] = [K_1] + [A_1]'[H_1] = [K_1] - [H_1]'[G]^{-1}[H_1]$$

$$[S_{zz}^{22}] = [K_2] + [A_2]'[H_2] = [K_2] - [H_2]'[G]^{-1}[H_2]$$

and

$$[S_{zz}^{gg}] = [K_g] + [A_g]'[H_g] = [K_g] - [H_g]'[G]^{-1}[H_g]$$

Finally, the SPR estimators of μ_1 = average volume per unit area on the first occasion, μ_2 = average volume per unit area on the second occasion, and μ_g = average change in volume per unit area from the first to the second occasion are respectively defined as

$$w_1 = [b]'[z^1], w_2 = [b]'[z^2], \text{ and } [w_g] = [b]'[z^g]$$

Their variances are estimated by the formulae

$$S_{w_1 w_1} = [b]'[S_{zz}^{11}][b] + [z^1]'[S_{bb}][z^1]$$

$$S_{w_2 w_2} = [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2]$$

and

$$S_{w_g w_g} = [b]'[S_{zz}^{gg}][b] + [z^g]'[S_{bb}][z^g]$$

Note that the new SPR estimators are generally different than the old SPR estimators of the previous section. This is because different sets of auxiliary variables are being used. However, the difference between the old and new estimators is expected to be small and probably not significantly different than zero from a statistical point of view.

Illustrative Numerical Example

Consider the sample data from $u = 117$ temporary plots measured in 1960, $m = 118$ permanent plots measured on two occasions in 1960 and 1964, and $n = 92$ temporary plots measured only in 1964. These plots are one-tenth acre in size and have been selected from a forest area in Québec, Canada

by a method which can be assumed to be equivalent to simple random sampling.

To the trees of these plots we shall apply a biomass regression function used by Cunia (1986a,b). For the sample data and the calculation of the regression function

$$u = b_1 + b_2 d + b_3 d^2$$

where u = tree biomass (total above ground green weight in pounds) and d = tree diameter (inches), the reader is referred to Cunia (1986c). We need here only the statistics

$$[b] = \begin{bmatrix} 5.1818118 \\ -25.653078 \\ 12.988357 \end{bmatrix} \quad \text{and}$$

$$[S_{bb}] = \begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

Because (i) the independent variables of the biomass regression function are 1 , d and d^2 , and (ii) the plot size is one-tenth of an acre, the plot variables s_1 , s_2 , and s_3 are defined as

$$s_1 = \Sigma(1/.10 \text{ acres})$$

$$= \text{number of trees per acre}$$

$$s_2 = \Sigma(d/.10 \text{ acres})$$

$$= \text{sum of tree diameters per acre, and}$$

$$s_3 = \Sigma(d^2/.10 \text{ acres})$$

$$= \text{sum of squared tree diameters per acre}$$

The values of s_1 , s_2 , and s_3 as measured in 1960 and/or 1964 on the $(u+m+n) = 327$ plots are given in a set of lecture notes by Cunia (1986d). We shall now use these values to calculate three sets of SPR estimators.

First Set: The New SPR Estimators -

To calculate these estimators, we shall need the following summary statistics calculated from the basic plot values s_1 , s_2 , and s_3

$$[\bar{s}_u] = \begin{bmatrix} 214.78632 \\ 1338.3291 \\ 10722.382 \end{bmatrix} = \text{vector of the 1960 averages calculated from the data of the } u = 117 \text{ plots}$$

$$[\bar{s}_n] = \begin{bmatrix} 228.69565 \\ 1502.8380 \\ 12521.600 \end{bmatrix} = \text{vector of the 1964 averages calculated from the data of the } n = 92 \text{ plots}$$

$$[\bar{s}_m^1] = \begin{bmatrix} 236.35593 \\ 1439.3161 \\ 10898.164 \end{bmatrix} = \text{vector of the 1960 averages calculated from the data of the } m = 118 \text{ plots}$$

$$[\bar{s}_m^2] = \begin{bmatrix} 266.01695 \\ 1638.6805 \\ 12320.176 \end{bmatrix} = \text{vector of the 1964 averages calculated from the data of the } m = 118 \text{ plots}$$

$$[S_{ss}^{11}] = \begin{bmatrix} 21272.932 & 110452.37 & 602351.30 \\ 110452.37 & 659011.32 & 4724541.6 \\ 602351.30 & 4724541.6 & 48604076 \end{bmatrix}$$

= estimate of the covariance matrix of the first measurement values s_1 , s_2 , and s_3 as calculated from the $m = 118$ plots

$$[S_{ss}^{12}] = \begin{bmatrix} 21987.926 & 118295.97 & 659678.73 \\ 109292.82 & 670898.34 & 4832348.2 \\ 542577.80 & 4443890.0 & 46552860 \end{bmatrix}$$

= estimate of the covariance matrix of the first measurement values s_1 , s_2 , and s_3 with the second measurement values s_1 , s_2 , and s_3 as calculated from the $m = 118$ plots

$$[S_{ss}^{22}] = \begin{bmatrix} 24039.555 & 123557.78 & 635935.23 \\ 123557.78 & 718316.31 & 4802637.7 \\ 635935.23 & 4802637.7 & 47000205 \end{bmatrix}$$

= estimate of the covariance matrix of the second measurement values s_1 , s_2 , and s_3 as calculated from the $m = 118$ plots

Note that (i) we have used superscripts to denote the first or second measurement (do not interpret them as exponents or powers), and (ii) the covariance terms of $[S_{ss}^{12}]$ related to the first measurement variables are arranged in rows and those of the second in columns. For example, the covariance of the first measurement s_1 with the second measurement s_2 is 118295.97 while the covariance of s_1 of the second measurement with s_2 of the first measurement is 109292.82.

By using now the formulae of the new SPR estimators, one can easily calculate and write down the vectors

$$[Q_1], [Q_2], [Q_g], \text{ and } [P].$$

and their covariance matrices

$$[K_1], [K_2], [K_g], [G], [H_1], [H_2], \text{ and } [H_g]$$

These are not given here, but the interested reader can find their values listed in Cunia (1986d). The inverse of $[G]$ is also listed there as well as the calculation of the following vectors of SPR coefficients.

$$[A_1] = \begin{bmatrix} -.57582423 & .27162431 & 2.1942318 \\ -.016295496 & -.71115549 & -.60394450 \\ .00082660796 & .0038778696 & -.60368650 \\ -.052395398 & -1.4196676 & -6.1216962 \\ .075951873 & .58362944 & 1.2028896 \\ -.0031360528 & -.011748779 & .23558038 \end{bmatrix}$$

$$[A_2] = \begin{bmatrix} .58503065 & 1.0689709 & 1.9045323 \\ -.051364277 & .14848762 & .068781280 \\ .0017007641 & .0047393886 & .30871816 \\ -.62752621 & .39234747 & 2.7713606 \\ -.021883489 & -.81539047 & -.73733567 \\ .0011022853 & .0052754490 & -.67282885 \end{bmatrix}$$

and $[A_g] = [A_2] - [A_1]$ not shown here.

Consequently the SPR estimates of

$$[\mu_z^1] = \begin{bmatrix} \text{mean number of trees per acre in 1960} \\ \text{mean sum of diameters per acre in 1960} \\ \text{mean sum of squared diameters per acre in 1960} \end{bmatrix}$$

$[\mu_z^2]$ similarly defined for 1964 measurement, and

$$[\mu_z^g] = [\mu_z^2] - [\mu_z^1]$$

are the following

$$[z^1] = \begin{bmatrix} 219.71326 \\ 1374.9418 \\ 10954.679 \end{bmatrix}, [z^2] = \begin{bmatrix} 247.57976 \\ 1561.1367 \\ 12280.514 \end{bmatrix},$$

$$\text{and } [z^g] = \begin{bmatrix} 27.866502 \\ 186.19492 \\ 1325.8344 \end{bmatrix}$$

and the estimates of their covariances are

$$[S_{zz}^{11}] = \begin{bmatrix} 65.995681 & 342.03116 & 1869.1504 \\ 342.03116 & 2039.9531 & 14673.089 \\ 1869.1504 & 14673.089 & 151545.20 \end{bmatrix}$$

$$[S_{zz}^{22}] = \begin{bmatrix} 75.556641 & 386.92007 & 1995.4171 \\ 386.92007 & 2243.7140 & 15044.211 \\ 1995.4171 & 15044.211 & 147808.27 \end{bmatrix}$$

and

$$[S_{zz}^{gg}] = \begin{bmatrix} 8.2793523 & 38.357716 & 219.65978 \\ 38.357716 & 209.87519 & 1597.4279 \\ 219.65978 & 1597.4279 & 17693.325 \end{bmatrix}$$

Consequently, the SPR estimators of μ_1 , μ_2 , and μ_g are calculated respectively as

$$w_1 = [b]'[z^1] = 108150.31 \text{ pounds}$$

$$w_2 = [b]'[z^2] = 120738.64 \text{ pounds}$$

$$w_g = [b]'[z^g] = 12588.336 \text{ pounds}$$

and their variances are estimated as

$$S_{w_1 w_1} = [b]'[S_{zz}^{11}][b] + [z^1]'[S_{bb}][z^1] = 17292284 + 11049512 = 28341796$$

$$S_{w_2 w_2} = [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2] = 16553974 + 11720724 = 28274698$$

and

$$S_{w_g w_g} = [b]'[S_{zz}^{gg}][b] + [z^g]'[S_{bb}][z^g]$$

$$= 2078027 + 43969 = 2121996$$

Note that the part of the total variance due to the error of biomass regression function is about 38.98, 41.45 and 2.07 percent respectively for w_1 , w_2 , and w_g . It may be interesting to compare these percentages with those obtained by Cunia (1986b) when the CFI estimators were calculated from the $m = 118$ permanent sample plots alone. When estimating the current average biomass per acre (w_1 and w_2) the error component due to the biomass regression is about the same in absolute terms (11049512 versus 10710584 for w_1 and 11720724 versus 11393835 for w_2) but about twice as large in relative terms (38.98 versus 18.58 for w_1 and 41.45 versus 20.37 for w_2). This is because the part of the total variance of the SPR estimators due to the error of the sample plots is smaller in absolute terms (17292284 versus 46923716 for w_1 and 16553974 versus 44529880 for w_2) than that of the CFI estimators; the SPR estimators are based on data from 353 plots while the CFI estimators are based on the data of the 118 permanent plots only. On the other hand, the error component of the variance of w_g due to the error of the biomass tables remains about the same for the SPR and CFI estimators, (43969 versus 50666 for the absolute and 2.07 versus 2.08 percent for the relative terms).

Second Set: The Common (Old) SPR Estimators

We thought it would be interesting to calculate the values (estimators and errors) by the common SPR procedure and compare them to the values of the first set. If we define the plot variables as

x_1 = plot biomass per acre at the first, 1960 measurement time, and

y_1 = plot biomass per acre at the second, 1964 measurement time we can proceed as follows.

We start with the calculations of individual plot values x_1 and y_1 . The ordinary procedure is to calculate the biomass of each individual tree separately, sum them up by sample plots and divide these sums by .10 acres, the area of the plot. Because the individual plot variables s_1 , s_2 , and s_3 are available, it is easier to use the formulae $x_1 = [b]'[s^1]$ and $y_1 = [b]'[s^2]$. Cunia (1986d) lists the individual plot values x_1 and y_1 and calculates the following basic statistics that are needed for the calculation of the second set of the SPR estimators

$$\bar{x}_u = 106046.85 = [b]'[\bar{s}_u]$$

$$\bar{y}_n = 125267.64 = [b]'[\bar{s}_n]$$

$$\bar{x}_m = 105851.11 = [b]'[\bar{s}_m^1]$$

$$\bar{y}_m = 119360.09 = [b]'[\bar{s}_m^2]$$

$$S_{xx} = 5536998459, S_{xy} = 5255342828,$$

$$S_{yy} = 5254525850$$

Note that there is only one variable (x and y) measured on the plot on a given occasion, not $p > 1$ and $q > 1$ as implicitly assumed in our formulae. This would simplify the calculation of the SPR estimators. Also, the statistics S_{xx} , S_{xy} , and S_{yy} were calculated only from the data of the $m = 118$ permanent plots.

Using the appropriate formulae, the reader can verify that

$$[Q] = \begin{bmatrix} \bar{x}_u \\ \bar{y}_n \\ \bar{y}_n - \bar{x}_u \end{bmatrix} = \begin{bmatrix} 106046.85 \\ 125267.64 \\ 19220.790 \end{bmatrix}$$

$$[P] = \begin{bmatrix} \bar{x}_u - \bar{x}_m \\ \bar{y}_n - \bar{y}_m \end{bmatrix} = \begin{bmatrix} 195.73999 \\ 5907.5513 \end{bmatrix}$$

$$[K] = \begin{bmatrix} 47324773 & 0 & -47324773 \\ 0 & 57114411 & 57114411 \\ -47324773 & 57114411 & 104439185 \end{bmatrix}$$

$$[G] = \begin{bmatrix} 94248489 & 44536804 \\ 44536804 & 101644292 \end{bmatrix}$$

$$[G]^{-1} = (10^{-9}) \begin{bmatrix} 13.380774 & -5.8629647 \\ -5.8629647 & 12.407167 \end{bmatrix}$$

$$[H] = \begin{bmatrix} 47324773 & 0 & -47324773 \\ 0 & 57114411 & 57114411 \end{bmatrix}$$

$$[A] = \begin{bmatrix} -.63324208 & .33485978 & .96810186 \\ .27746347 & -.70862804 & -.98609152 \end{bmatrix}$$

= matrix of SPR coefficients

Consequently, the (common) SPR estimates are

$$[w] = \begin{bmatrix} w_1 \\ w_2 \\ w_g \end{bmatrix} = [Q] + [A]'[P] = \begin{bmatrix} 107562.03 \\ 121146.93 \\ 13584.900 \end{bmatrix}$$

and

$$[S_{ww}] = [K] + [A]'[H] =$$

$$= \begin{bmatrix} 17356735 & 15847163 & -1509572 \\ 15847163 & 16641538 & 794375 \\ -1509572 & 794375 & 2303947 \end{bmatrix}$$

Note that one can verify that

$$S_{w_g w_g} = S_{w_1 w_1} - 2S_{w_1 w_2} + S_{w_2 w_2}$$

$$= 17356735 - (2) (15847163) + 16641538 = 2303947$$

It may be interesting to compare these results to those obtained before as the first set. As one can easily verify the results are not identical. This is not surprising, since the two SPR procedures do not use the same auxiliary variables. The second set is based only on the biomass per acre variables of the first and second measurement, while the first set is based on plot variables s_1 , s_2 , and s_3 .

The differences between the two sets of measurements may look, at first sight, large. If we consider first the difference between the estimates w_1 and w_2 , we find the values

$$108150.31 - 107562.03 = 588.28 \text{ or about } .54 \text{ percent}$$

and

$$120738.64 - 121146.93 = -408.29 \text{ or about } .34 \text{ percent}$$

Obviously these differences are negligibly small, about 11.05 and 7.68 percent of the standard error of w_1 and w_2 respectively. The difference is, however, much larger for the estimate w_g since

$$12588.336 - 13584.900 = -996.564 \text{ or about } 7.9 \text{ percent}$$

But this is still only about 68.41 percent of the standard error of w_g of the second set of estimators.

The variances of the second set of SPR estimators are much smaller than those of the first set. This is not surprising since the error of the second set does not include the error of the biomass regression function. When this last error is taken out from the error of the first set, the estimates of the variances become extremely close. The slightly smaller values of the first set are due to the fact that the first set uses three auxiliary variables s_1 , s_2 , s_3 , while the second set uses only a single variable, the biomass per acre, that is nothing but a linear combination of s_1 , s_2 and s_3 , say x or y equal to $b_1s_1 + b_2s_2 + b_3s_3$.

To see what happens when the common (old) SPR estimators are calculated with s_1 , s_2 , and s_3 used as the auxiliary variables, we have calculated, as an example the following additional SPR estimator of μ_2 .

Third (Old) SPR Estimator of μ_2 : s_1 , s_2 , and s_3 Used as Auxiliary Variables

To calculate this estimator, we proceed as follows:

$$[Q] = [\bar{y}_n] = [125267.64]$$

$$[P] = \begin{bmatrix} [\bar{s}_u] - [\bar{s}_m] \\ \bar{y}_n - \bar{y}_m \end{bmatrix} = \begin{bmatrix} -21.569607 \\ -100.98704 \\ -175.78196 \\ 5907.5513 \end{bmatrix}$$

$$[K] = [S_{yy}/n] = [57114411]$$

$$[G] = \begin{bmatrix} \left(\frac{u+m}{um}\right) [S_{ss}^{11}] & \left(\frac{1}{m}\right) [S_{sy}^{12}] \\ \left(\frac{1}{m}\right) [S_{sy}^{12}]' & \left(\frac{m+n}{mn}\right) [S_{yy}] \end{bmatrix}$$

$$= \begin{bmatrix} 362.09902 & 1880.0743 & 10252.974 & 47859.527 \\ 1880.0743 & 11217.417 & 80419.186 & 390847.37 \\ 10252.974 & 80419.186 & 827318.40 & 4181841.0 \\ 47859.527 & 390847.37 & 4181841.0 & 101644292 \end{bmatrix}$$

$$[H] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ S_{yy}/n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 57114411 \end{bmatrix}$$

Note that to calculate $[G]$ we need the additional covariance terms of s_1 , s_2 , and s_3 (of the first measurement) with y (the biomass per acre of the second measurement). The values of these covariances divided by $m = 118$ are shown in $[G]$ as 47859.527, 390847.37, and 4181841.0.

The reader can verify that

$$[A] = [G]^{-1}[H] = \begin{bmatrix} 12.787013 \\ -6.5898007 \\ 4.0713794 \\ -7.1009051 \end{bmatrix}$$

$$w_2 = [Q] + [A]'[P] = 120746.74$$

and

$$S_{w_2 w_2} = 16558010$$

Note that the newly obtained values (of the old SPR estimators) of 120746.74 and 16558010 are much closer to the corresponding values (of the new SPR estimators) of 120738.64 and 16553974.

Concluding Remarks

The approach used by Cunia (1986a) to combine the error of volume tables with that of sample plots when calculating volume and growth estimates in Continuous Forest Inventory (CFI) systems is extended to the case where the Sampling with Partial Replacement (SPR) is used to increase the efficiency. We have considered the simplest case, that of forest inventories on two successive occasions. By properly defining the cluster variables s_1 , s_2 , ..., one can apply the same formulae to clusters of both types, sample plots of fixed area or relascope sample points.

The new SPR estimators are slightly different (although, for our sample data not significantly different) than the classical ones where the error of the biomass tables is being ignored. This is due to the fact that the classical SPR method is applied to linear combinations of cluster variables s_1 , s_2 , ... (namely cluster biomass per unit area), while the new SPR approach is applied directly to

these variables. In this sense the new SPR estimators would normally be better; SPR regressions on linear combinations of auxiliary variables would not generally be better than regression on the variables themselves.

We have considered only one species. By defining the giant size vector $[B]$ of regression coefficients

$$[B]' = [[b^1]' [b^2]' \dots [b^c]']$$

where $[b^i]$ is the vector of regression coefficients of species $i = 1, 2, \dots, c$, and its giant size covariance matrix $[S_{BB}]$ whose ij -th submatrix component is the covariance matrix $[S_{ij}^{bb}]$ of $[b^i]$ and $[b^j]$, one can extend the methodology from one to more than one species. This requires the definition of a giant size vector of cluster (plot or relascope point) values

$$[s]' = [[s^1]' [s^2]' \dots [s^c]']$$

where $[s^i]$ is the subvector of cluster values corresponding to species i , and the giant size covariance matrix $[S_{ss}]$ of $[s]$ whose ij -th submatrix component is the covariance matrix $[S_{ij}^{ss}]$ of $[s^i]$ and $[s^j]$. From here on, the procedure and the formulae introduced here apply immediately; the giant size vector $[Z]$ is defined as

$$[Z]' = [[z^1]' [z^2]' \dots [z^c]']$$

where $[z^i] = [\bar{s}^i] =$ the sample average of $[s^i]$, and the giant size covariance matrix $[S_{zz}]$ contains as submatrix component $[S_{ij}^{zz}] = [S_{ij}^{ss}]/n$.

The approach can also be generalized to more than two measurements by using the SPR approach outlined by Cunia and Chevrou (1969) and Newton, Cunia and Bickford (1974). It suffices to define the vectors $[Q]$ and $[P]$ so as to (i) include the averages from all measurements on all auxiliary variables, and (ii) have the SPR estimators $w = [Q] + [A]'[P]$ unbiased.

Acknowledgements

The 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. Continuous forest inventory, partial replacement of samples and multiple regression. Forest Science, 11: 480-502; 1965.

Cunia, T. The error of forest inventory estimates: its major components. In: Proceedings of the workshop on "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; 1986a.

Cunia, T. On the error of forest inventory estimates: Continuous Forest Inventory without SPR. In: Proceedings of the workshop on "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; 1986b.

Cunia, T. Construction of the tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "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; 1986c.

Cunia, T. On the error of biomass estimates in forest inventories: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (86-001), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

Cunia, T.; Chevrou, R.B. Sampling with partial replacement on three or more occasions. Forest Science 15: 204-224; 1969.

Jacobs, M.W.; Cunia, T. Use of dummy variables to harmonize tree biomass tables. Canadian Journal of Forest Research, 10:483-490; 1980.

Newton, C.M.; Cunia, T.; Bickford, C.A. Multivariate estimators for sampling with partial replacement on two occasions. Forest Science, 20:106-116; 1974.

Schreuder, H.T.; Swank, W.T. A comparison of several statistical models in forest biomass and surface area estimation. In: Forest Biomass Studies. H.E. Young (Ed.), University of Maine at Orono, Maine; 1971.

Ware, K.D.; Cunia, T. Continuous forest inventory with partial replacement of samples. Forest Science Monograph 3; 1962.

RESEARCH PAPERS

**Biomass Regressions
And Measurement Error**

Moderator: Timothy G. Gregoire

AN OPTIMIZATION MODEL FOR SUBSAMPLING TREES FOR
BIOMASS MEASUREMENT

Tiberius Cunia

Professor of Statistics and Operations Research
SUNY College of Environmental Science & Forestry
Syracuse, NY 13210

When selecting sample trees for biomass tables construction is customary to estimate their biomass by subsampling rather than measure their actual values. By this procedure one must work with estimated rather than actual tree biomass. One may use an intensive tree subsampling and obtain estimates of the sample tree biomass with high precision. But this implies that for the same sampling costs one must reduce the number of sample trees. On the other hand one may wish to work with more sample trees and less precise estimates of their biomass. An approach using mathematical programming models is shown that optimizes the combination of number of sample trees and intensity of subsampling.

Introduction

Although not explicitly stated when they are applied to forest inventory problems, the tree biomass regression functions are not without error. This error has four main components, each component being associated with a specific source.

There is first the error due to sample tree selection. The same sampling method applied on different occasions results in different sets of sample trees and, thus, in different biomass regression functions. The size of this error component, generally known as the sampling error, is affected by the sampling design by which the sample trees are selected, by the number of sample trees, and by the inherent variation between the biomass of various trees.

Once the sample trees are selected, they must be measured for their biomass. The same tree measured on different occasions may result in different biomass values. We hypothesize the existence of a constant, fixed value which we shall call true biomass of a tree (or of some of its components). What we call the biomass of a tree is the measured value that we have obtained by some measurement or subsampling process. The difference between the true and the measured biomass value will be known here as the measurement error. This is the second error component. For variables that can be measured directly, as for example, the green weight of some small tree component, the measurement error may be negligibly small. For others, like the oven-dry weight of the entire tree which, for practical purposes is measured by subsampling, the error may be quite large.

The third source of error is the statistical model used in data analysis and estimation of the biomass regression function. Different statisticians working with the same sample data may arrive at different biomass regressions. For any given real world problem, one can construct different models, each model having its own assumptions. These assumptions refer to the characteristics of the tree population and the method by which the sample trees are selected. If valid statistical inferences are desired, the assumptions must be sufficiently well satisfied. But even if they are satisfied, different, equally good statistical models would generally yield different biomass regressions. If the models are well selected, however, the difference between their results are expected to be small. This component is known as the error due to statistical modeling. It should not be confused with what is known as statistical error, a term used to denote, in general, the error (from all but subjective sources) associated with the difference between our inferences and the true values of the parameters we want estimated.

The fourth error component is due to the application of the biomass regression function to a specific case and may be known as the error due to application. This is because the regressions are estimated from trees that are no longer members of the population from which they were drawn; they were felled down and destroyed. Furthermore, the populations of trees are dynamic (they change with time) and it is quite common to apply biomass regression functions to forest areas other than those for which the regression functions were calculated. This is generally known as a non-statistical error component since it can hardly be evaluated by statistical means. Note that strictly speaking, this error component is not a part of the error of the biomass regression; only a component due to its use.

We shall not be concerned here with the first, the third or the fourth error component. We shall assume that (i) the sample trees are selected by simple random sampling, the usual assumption of the least squares method, (ii) the statistical model used in the data analysis is appropriate, and (iii) the biomass regression function is being applied to the population for which it was calculated. We shall only be concerned here with the second component, the measurement error.

The measurement error itself has three main subcomponents. There is first the so-called measurement bias due to faulty instruments, to poorly defined or applied measurement procedures, to poor subsampling designs or to selection of biased estimators. There is also the random error due to the precision of the measuring devices or the measurement process itself, as for example, the precision of the weighing scale, caliper or diameter tape, etc. This last error can be assumed to be negligibly small, and together with the measurement bias above will be ignored in this discussion. Our only topic of discussion will be the subcomponent error due to the inherent variation associated with any subsampling process used to determine the biomass (usually oven-dry weight of large tree components) of a sample tree. The same subsampling

procedure, using the same estimators, applied on various occasions, would normally yield different results.

Note that there are no statistical estimation problems related to the subsampling error. As long as (i) the expected value of this error is zero, that is, the estimator of the biomass of the sample tree is unbiased, and (ii) the error is statistically independent of the error component due to sample tree selection, the least squares method takes it automatically into account. The only problem is one of efficiency. Should one select a small number of sample trees and estimate their biomass by an intensive subsampling or is it better to work with a large number of sample trees for which the biomass is poorly measured?

The specific objectives of the present study are to construct a mathematical model for the error of tree biomass regression functions. Because the same error can be obtained from various combinations of number of sample trees and intensity of subsampling, a model is needed to guide us in the selection of the sample and subsample size. Optimum sample and subsample size should be sought; those that would minimize the costs of sampling for desired precision of tree biomass regressions or those that maximize the precision for given allowable sampling costs.

The type of model we shall construct is known as a mathematical programming model. It will consist of either (i) an objective cost function that should be minimized subject to the side condition that the error of the biomass regression function should be smaller than a predetermined value, or (ii) an objective error function that should be minimized subject to sampling costs being below a specified amount. Both cost and error functions will be expressed explicitly in terms of sample size n (of sample trees) and subsample size m (for the estimation of the biomass of the sample trees).

In constructing this model, we shall assume that the selection of sample trees and the estimation of their biomass by subsampling is made by the simple random sampling method. We shall then extend the model to include the case where the biomass of the sample trees is estimated by other sampling methods. We shall not consider, however, the case where the sample trees themselves are selected by methods other than simple random sampling. This is because in order to calculate the regression function, one must then modify the least squares method of estimation.

Expressing the Error of Biomass Regressions

The error of the biomass regression functions can be expressed in many ways, each way being adequate for some applications and quite inadequate for others. It can be expressed as the variance or standard error of (i) regression estimates of the biomass of trees of given size, (ii) regression coefficients, (iii) estimators (of parameters of interest) obtained when the biomass regressions

are applied to forest inventory data, etc. It can also be expressed as confidence intervals (at some predetermined probability level) of (i) the expected biomass of trees of given size, (ii) parameters of interest of the true regression function, (iii) parameters of interest of some forest tree population, etc.

We are interested in the application of the biomass regressions to forest inventory data and wish to work with an expression of the error that can be used within this context. Cunia (1986a, b) has suggested a method to combine the error of the biomass regression with the error of the sample plots (of the forest inventory) so that the calculated error of the inventory estimates would contain the error from both sources. This method assumes that (i) the true regression function of tree biomass on diameter, height, species, etc. is linear, that is, of the form

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p = [\beta]'[x]$$

where y = tree biomass, $x_i = 1, x_i, i = 2, 3, \dots, p$ are tree characteristics other than biomass and $[]$ denotes a vector or matrix (with $[]'$ denoting a transpose), (ii) the sample plots (where the trees are measured for attributes other than biomass) are selected by a variety of sampling designs, (iii) the sample of plots and the biomass regression are statistically independent, and (iv) the forest parameters of interest are average biomass per tree, average biomass per unit area (acre or hectare), average growth (or growth components of biomass per unit area, etc.).

To apply Cunia's procedure, one must have an estimate $[b]$ of the vector $[\beta]$ of regression coefficients and an estimate $[S_{bb}]$ of the covariance matrix $[\sigma_{bb}]$ of $[b]$. Because $[\sigma_{bb}]$ can be expressed as a function of (i) the conditional variance of y given $[x]$ and (ii) the sample sizes n and m defined in the previous section, the error expression we shall select for our optimization model is that of $[\sigma_{bb}]$.

To better see our approach, let us review the least squares method as applied to biomass regressions. We shall use the following basic assumptions.

(1) The true regression function of y on $x_1 = 1, x_2, x_3, \dots, x_p$ is of the linear form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p = [\beta]'[x]$$

(2) The conditional variance of y given $[x]$, say $\sigma_{yy|x}$, is proportional to some given function of the known tree variables $x_1, x_2, \dots, x_p$, say

$$\sigma_{yy|x} = \sigma_{uu|v}^w$$

where $\sigma_{uu|v}$ is the unknown factor of proportionality and w is a known function of $x_1, x_2, \dots, x_p$.

(3) The n sets of sample observations $y_k, x_{k1}, x_{k2}, \dots, x_{kp}$ for $k = 1, 2, \dots, n$ are statistically independent of each other.

Then, the best linear and unbiased estimator $[b]$ of $[\beta]$ is the weighted least squares estimator calculated as follows.

Step 1 - Calculate the n sets of transformed sample values $u_k = y_k / \sqrt{w_k}, v_{ki} = x_{ki} / \sqrt{w_k}, i=1,2,\dots,p$ and $k = 1, 2, \dots, n$ and construct the vector $[U]$ and matrix $[V]$ of these values.

$$[U] = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \text{ and } [V] = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{np} \end{bmatrix}$$

Step 2 - Calculate the matrices [T] and [P] of sum of cross-products

$$[T] = [V]'[V] \text{ and } [P] = [V]'[U]$$

and the estimator [b] of [β]

$$[b] = [T]^{-1}[P]$$

Step 3 - Calculate the estimator $S_{uu|v}$ of $\sigma_{uu|v}$

$$S_{uu|v} = ([U]'[U] - [b]'[P])/(n-p)$$

and the estimator $[S_{bb}]$ of the covariance matrix $[\sigma_{bb}]$ of [b]

$$[S_{bb}] = S_{uu|v}[T]^{-1}$$

Remarks

(1) If y is the biomass of the main tree bole or the total above ground tree biomass, it has been found empirically that the conditional variance of y given diameter d or given diameter d and height h is approximately proportional to d^4 or d^4h^2 respectively. This means that for the first case u is equal to y/d^2 while u is equal to y/d^2h for the second case. Similar definitions apply to variables $v_1, v_2, \dots, v_p$. Of course, for other biomass components the conditional variance of y given d or given d and h may be proportional to known values other than d^4 or d^4h^2 , may be related in a different way to d or d and h or may even be homogeneous.

(2) $S_{uu|v}$ is an estimator of the factor of proportionality $\sigma_{uu|v}$ which now can also be interpreted as the conditional variance of $u = y/\sqrt{w}$ for given [v], or what is the same thing, for given [x]. This means that the conditional variance of y given [x] can be estimated by the formula

$$S_{yy|x} = w S_{uu|v}$$

For the biomass y of (1) above, we have

$$S_{yy|x} = d^4 S_{uu|v}$$

$$S_{yy|x} = d^4 h^2 S_{uu|v}$$

(3) If the model assumptions stated above are satisfied, the conditional variance of u given [x] is homogeneous, that is $\sigma_{uu|v}$ is the same for the trees of all sizes [x]. This has an extremely important implication. As we shall analyze the conditional variance of y given [x] which varies with the tree size [x], and because $\sigma_{uu|v} = \sigma_{yy|x}/w$ is constant over the trees of all sizes [x], we can work with the variance u and then interpret the conclusions reached later in terms of the variable y.

For example, if the variable x are all defined in terms of d alone, and the conditional variance of y given d is proportional to d^4 , the variable $u = y/d^2$ has a conditional variance which is

homogeneous with respect to d; it is the same for the trees of any diameter size d. This means that if the trees are classified by diameter d, then the variance of u within a diameter class is the same for all diameter classes. Note that $u = y/d^2$ has a physical meaning; it represents "biomass per square inch of breast high diameter." As the basal area of a tree is $\pi d^2/4$, and $\pi/4$ is a constant value, u is equivalent to the variable $u^* = (y/\text{basal area})$, that is equivalent to the variable tree biomass per unit of basal area.

Let us express now the covariance matrix $[\sigma_{bb}]$ of [b], defined as

$$[\sigma_{bb}] = \sigma_{uu|v}[T]^{-1}$$

as an explicit function of the number n of sample trees. If the n sample trees are selected completely at random, it can be shown that the expected value of [T] can be written as

$$[T] = n[\bar{T}]$$

where the ij-th element of $[\bar{T}]$ is the expected value of the cross product $v_i v_j$. As the inverse of [T] is

$$[T]^{-1} = [\bar{T}]^{-1}/n = [\theta]/n$$

where $[\theta]$ is the inverse of the matrix $[\bar{T}]$, we can finally write that, on the average, we expect to have the approximate relationship

$$[\sigma_{bb}] = (\sigma_{uu|v}/n)[\theta]$$

A Simple Optimization Model 1

Consider a sample of n trees selected by simple random sampling from a relatively large population of N trees. If y is the true biomass of a tree (or some of its components), we shall assume that $\sigma_{yy|x}$, the conditional variance of y given [x] is proportional to some known value w (usually, but not necessarily a function of the known tree values $x_1, x_2, \dots, x_p$). As shown in the previous section, the variable $u = y/\sqrt{w}$ has a conditional variance for given [x] that remains constant over all values [x]. We shall denote this variance V_u . Note that V_u is defined in terms of true values y and $u = y/\sqrt{w}$.

Because the biomass value y of a given tree is not generally measured but estimated by subsampling, let us further assume that a given tree is subdivided into a large number M of approximately equal parts denoted here as biomass samples. These samples are usually, but not necessarily non-overlapping. Assuming non-overlapping biomass samples, let m samples be selected by simple random sampling. If

$$z_k = (\text{biomass of sample } k)/\sqrt{w} = \text{variable measured (without error) on the biomass sample } k$$

$$\mu_z = \Sigma z_k/M = u/M = \text{mean of } z \text{ within a given tree}$$

$$\sigma_{zz} = \Sigma (z_k - \mu_z)^2/M = \text{variance of } z \text{ within a given tree}$$

where Σ is taken over all samples $k = 1, 2, \dots, M$, then, it is well known that for Σ taken now over the subsample values $k = 1, 2, \dots, m$, we have

$$\bar{z} = \Sigma z_k/m = \text{subsample mean of } z = \text{estimator of } \mu_z$$

$$S_{zz} = \Sigma (z_k - \bar{z})^2/(m-1) = \text{subsample variance of } z = \text{estimator of the variance of } z$$

$\hat{u} = M\bar{z} =$ estimator of the true value of the biomass u of the given tree

$\sigma_{uu}^{\wedge} = M^2\sigma_{zz}/m =$ variance of $\hat{u}$ as an estimator of u of the given tree, and

$S_{uu}^{\wedge} = M^2S_{zz}/m =$ estimator of $\sigma_{uu}^{\wedge}$

Note that for the given tree, the biomass y is estimated by

$$\hat{y} = \hat{u}\sqrt{w} = M\bar{z}\sqrt{w}$$

and the variance of $\hat{y}$ is estimated by

$$S_{yy}^{\wedge} = wS_{uu}^{\wedge} = wM^2S_{zz}/m$$

Furthermore, under our subsampling assumptions, $\bar{z}$, $\hat{u}$, and $\hat{y}$ are all unbiased and when $\sigma_{uu}^{\wedge}$, $S_{uu}^{\wedge}$, and $S_{yy}^{\wedge}$ were defined, the effect of the finite population correction factor was ignored.

For convenience, we shall write $V_2 = M^2\sigma_{zz}$. This means that $\sigma_{uu}^{\wedge}$ can be written as V_2/m . Recall that V_1 has been defined as the conditional variance of u given $[x]$. As u is not known, we have to work with its estimator $\hat{u}$. This implies that the variance of $\hat{u}$ about the conditional mean of u given $[x]$ has two, statistically independent components; the first component is V_1 associated with the variation of u about its conditional mean and the second component is V_2/m associated with the variation of $\hat{u}$ about the true value u . Consequently, the conditional variance of u (which in reality is $\hat{u}$) given $[x]$, which in the previous section was denoted by $\sigma_{uu|v}$, can now be written as

$$V = V_1 + V_2/m = \sigma_{uu|v}$$

As an example of a case where this model can be applied, consider the biomass y of the main bole of randomly selected trees and its regression function on diameter d . Empirical evidence suggests that the regression function of y on d is of the approximate form

$$y = \beta_1 + \beta_2 d + \beta_3 d^2 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = [\beta]'[x]$$

and that the conditional variance of y given $[x]$ is approximately proportional to $w = d^4$. Consequently, the variable u is defined as

$$u = y/d^2 = \text{bole biomass per square inch (of squared diameter)}$$

Note that y and u are true values of a given tree.

To estimate y and u (of a tree of known d), we shall divide the bole into a large number M of non-overlapping disks. If m disks are selected by simple random sampling without replacement, then we can define the following:

$z = (\text{disk biomass})/d^2 =$ value measured on each disk

$\mu_z = u/M =$ true mean of z

$\sigma_{zz} =$ variance of z

$\bar{z} =$ sample mean of the m biomass samples

$S_{zz} =$ sample variance of the m biomass samples

$\bar{z} =$ unbiased estimator of μ_z

$\hat{u} = M\bar{z} =$ unbiased estimator of u

$\sigma_{uu}^{\wedge} = M^2\sigma_{zz}/m = V_2/m =$ variance of $\hat{u}$ (as an estimator of u)

$S_{uu}^{\wedge} = M^2S_{zz}/m =$ estimator of $\sigma_{uu}^{\wedge}$

$\hat{y} = \hat{u}d^2 =$ unbiased estimator of y

$\sigma_{yy}^{\wedge} = \sigma_{uu}^{\wedge}d^4 =$ variance of $\hat{y}$ (as an estimator of y)

$S_{yy}^{\wedge} = S_{uu}^{\wedge}d^4 =$ estimator of $\sigma_{yy}^{\wedge}$

Because y and u are not known, we must work with $\hat{y}$ and $\hat{u}$. The conditional variance of $\hat{y}$ about the regression function of y on $[x]$ (that is, d) is assumed proportional to d^4 , while the conditional variance of $\hat{u}$ about its regression function of u on $[v]$ (that is, d) is homogeneous. Because $\hat{u}$ as an estimate of u is statistically independent of u as an estimate of the expected value of u given $[v]$ (or $[x]$ or d), the conditional variance of $\hat{u}$ given $[v]$ can be written as

$$\sigma_{uu|v} = V_1 + V_2/m$$

since the conditional variance of u given $[v]$ has been denoted by V_1 and the variance of $\hat{u}$ as an estimator of u has been written as V_2/m .

As a second example of where this model can be applied is the method of estimating the tree biomass y by the randomized branch method described by Valentine, Tritton, and Furnival (1985). Each application of the method yields an estimate $\hat{y}$ of y , and because d is known, an estimate $\hat{u} = \hat{y}/d^2$ of $u = y/d^2$. As one can apply the method infinitely many times, M is infinite in size. Nevertheless, we can define the following:

$z_k =$ estimate of u based on the k -th application of the randomized branch method

$\bar{z} = \hat{u} =$ estimate of u , since the expected value μ_z of z is equal to u (the method is unbiased)

$V_2/m = \sigma_{zz}/m =$ variance of $\bar{z} =$ variance of $\hat{u}$.

Note that with this subsampling procedure (i) the biomass samples are overlapping, (ii) the m biomass samples are selected with equal probabilities, even though various parts of the tree enter in biomass samples with unequal probabilities, (iii) each value z is defined so as to be an unbiased estimate of u and (iv) the variance of $\hat{u}$ about the regression function (of u on $[v]$) can be written as $V = (V_1 + V_2/m)$.

Let us define now the sampling costs components,

$c_s =$ average costs directly related to the selection and measurement of a biomass sample, and

$c_t =$ average costs of selecting, processing, and measuring a sample tree for characteristics other than those directly related to the measurement of the biomass samples.

For example, the costs of moving to and from the sample trees, the costs of measuring the tree diameter, height or bole length, the costs of felling, branching and measuring the green weight of the entire tree bole, the costs of the separation (conceptually if not physically) of the tree bole into non-overlapping disks, etc. are all included in c_t .

on the other hand, the actual selection of a sample disk or its location, the measurement of the green or oven-dry weight of a sample disk, etc. are all included in c_s .

If the fixed costs of sampling not directly related to selection and measurement of sample trees and biomass samples (that is, all costs not included in c_t and c_s) are ignored, the variable costs of sampling and subsampling a tree for its biomass can be evaluated, at least approximately, by the linear cost function

$$C = nc_t + nmc_s$$

As an example of some of the fixed sampling costs are those of training costs of the field crews, the costs of writing and debugging computer programs, etc.

It has been shown in the previous section that the expression of the error of biomass regression we shall work with is $[\sigma_{bb}] = (\sigma_{uu}/\sqrt{n})[0]$. If the selection of the sample tree is done by simple random sampling, it has been shown that $[0]$ is a parameter of the tree population of interest. Consequently, it suffices to work only with the expression $(\sigma_{uu}/\sqrt{n})$ which is a function of both n and m as the $\hat{u}$ variables are estimated by the $\hat{u}$ values determined by subsampling. Thus, the problem reduces to that of finding the sample sizes m and n for which the sampling costs C are minimized and the variance

$$\sigma_{uu}/\sqrt{n} = V/n = (V_1 + V_2/m)/n$$

is equal to some desired value K_0 . Or equivalently, the problem is that of finding m and n for which the value of the variance $K = V/n$ is minimized for given costs of sampling C_0 . More specifically, this optimization problem can be formalized as follows.

Optimization Problem 1. Find m and n that minimize the cost (objective) function

$$C = nc_t + nmc_s$$

subject to the side (variance) condition that

$$(V_1 + V_2/m)/n = K_0$$

where all parameters c_t , c_s , V_1 , V_2 , and K_0 are positive values and m and n are positive integers, and

Optimization Problem 2. Find m and n that minimize the variance (objective) function

$$K = (V_1 + V_2/m)/n$$

subject to the side (cost) condition that

$$nc_t + nmc_s = C_0$$

where c_t , c_s , V_1 , V_2 , and C_0 are all positive and m and n are positive integers.

By using calculus optimization techniques, one can easily find the following optimum solutions.

Solution to Problem 1. Nearest positive integer values

$$m = \sqrt{c_t V_2 / c_s V_1} = (\sqrt{c_t / c_s}) (\sqrt{V_2 / V_1})$$

and

$$n = (V_1 + V_2/m)/K_0 = (\sqrt{c_t V_1} + \sqrt{c_s V_2}) (\sqrt{V_1 / c_t}) / K_0$$

Solution to Problem 2. Nearest positive integer values

$$m = \sqrt{c_t V_2 / c_s V_1} = (\sqrt{c_t / c_s}) (\sqrt{V_2 / V_1})$$

and

$$n = C_0 (\sqrt{V_1 / c_t}) / (\sqrt{c_t V_1} + \sqrt{c_s V_2})$$

We shall show below the proof to the first problem; the proof to the second problem follows the same lines and can be easily done by the reader.

Proof of Solution to Problem 1. To find the optimum values m and n , one can use the method of Lagrange multipliers or the method of substitution. Using the latter method, we substitute the value of n from the side condition

$$n = (V_1 + V_2/m)/K_0$$

for the value of n of the objective function. Then, the new objective function to minimize becomes

$$C = (V_1 + V_2/m) (c_t + mc_s) / K_0$$

Because $K > 0$ and $(V_1 c_t + V_2 c_s) > 0$ are both fixed values, it is equivalent to work with the objective function

$$C^* = K_0 C - (c_t V_1 + c_s V_2) = (V_1 + V_2/m) (c_t + mc_s) - (c_t V_1 + c_s V_2) \\ = c_s V_1 m + c_t V_2 / m$$

To find the optimum solution, we equate to zero the first derivative of C^* with respect to m and solve for m the resulting equation. Then

$$dC^*/dm = c_s V_1 - c_t V_2 / m^2 = 0$$

and the only admissible solution (m is not allowed to be negative) is

$$m = \sqrt{c_t V_2 / c_s V_1}$$

As the second derivative

$$d^2 C^* / dm^2 = 2c_t V_2 / m^3$$

is always greater than zero for the positive solution value of m above, the function C^* attains a minimum. This in turn yields the optimum value of n as

$$n = (V_1 + V_2/m)/K_0 \\ = (V_1 + \sqrt{c_s V_1 V_2 / c_t}) / K_0 \\ = (\sqrt{V_1 / V_2} + \sqrt{c_s / c_t}) (\sqrt{V_1 V_2}) / K_0 \\ = (\sqrt{c_t V_1} + \sqrt{c_s V_2}) (\sqrt{V_1 / c_t}) / K_0$$

Remarks

(1) Because m and n can only be positive integers, the optimum solution above, which in general will be a non-integer solution, must be rounded-off to the nearest integer or to the nearest integer above. This is not a critical factor since the cost values c_t and c_s and variance values V_1 and V_2 are only approximately known.

(2) The optimum subsample size m does not depend on n but only on the variance components V_1 and V_2 and the average costs of sampling per tree and per biomass sample c_t and c_s respectively. Once the optimum m is determined, the value of n is then determined as a function of m and the desired precision K or allowable costs C . This is a very important finding. It means that we can optimize the subsampling first and then seek the value of the sample size n of trees to satisfy the overall requirements of precision or costs.

(3) At least from an intuitive point of view, the optimum solution makes sense. As the variance component V_2 (between biomass samples within trees) decreases with respect to the variance component V_1 (between true value of tree biomass), and as long as the costs c_t and c_s remain the same, we would normally expect m to decrease; we should spend relatively less money for subsampling. For the limiting case where $V_2 = 0$, we must have the limiting case $m = 1$, since this is the positive integer closest to solution $m = 0$; only one biomass sample would give us all the necessary information about the true value u . Similarly, as the costs c_s increase with respect to c_t and the variance components V_1 and V_2 remain the same, we would expect m to decrease; we should spend relatively less money for subsampling. For the limiting case where $c_s = \infty$, we have the limiting case $m = 1$, the positive integer closest to $m = 0$.

(4) We have assumed here implicitly, that the variance component V_2 is the same from tree to tree. This is obviously not true as each sample tree $i = 1, 2, \dots, n$ has its own variance V_{2i} . Because (i) the values V_{2i} are random variables (the trees are selected at random), (ii) knowledge about the probability distribution of V_{2i} is required, and (iii) to construct an optimization model where V_2 itself becomes a random variable is extremely difficult, we must work with some kind of average value V_{2i} , or an estimate of the expected value of V_{2i} , which here is denoted simply as V_2 . This same remark applies as well to the costs c_s and c_t which are also averages of random variables c_{si} and c_{ti} . In general c_{ti} increases with the tree size (felling and processing a large tree is more expensive than felling and processing a small tree), even though the variation of c_{si} (of the costs per biomass sample), for some tree components at least may be relatively small. For practical considerations we must work with the expected values c_t and c_s of the costs c_{ti} and c_{si} respectively.

(5) We have assumed here that a fixed number m of biomass samples are to be selected from each sample tree. A better optimization model may be that which makes m a function of V_2 (or diameter d , if V_2 is a function of d), that is, having an optimum solution m_i as a function of V_{2i} of the sample tree i we happen to select. This type of a model is obviously of much more complexity. However, due to the usually poor knowledge that one may have about the costs c_t and c_s or variance components V_1 and V_2 , it is doubtful that one may gain much by constructing such complex models; use of models based on average values V_2 , c_s and c_t are sufficiently good for all practical purposes.

Optimization Models 2, 3, 4, and 5

It is relatively easy to extend the usefulness of the simple optimization model 1 of the previous section by considering a few more common subsampling designs or estimators. For convenience these new models will be known as Models 2, 3, 4, and 5.

Model 2 - Simple Random Subsampling; m is large with respect to M . When m is too large with respect to M one cannot ignore the effect of the finite population correction factor. As the variance of $\hat{u}$ is equal to

$$\sigma_{\hat{u}\hat{u}} = \left(\frac{M-m}{M-1} \right) \left(\frac{M^2 \sigma_{zz}}{m} \right) \approx \left(\frac{M-m}{M} \right) \left(\frac{M^2 \sigma_{zz}}{m} \right) = (M-m) V_2 / Mm$$

we can write the conditional variance of $\hat{u}$ for given $[x]$ as

$$V = V_1 + (M-m) V_2 / Mm = V_1^* + V_2 / m$$

where $V_1^* = V_1 - V_2 / M$.

As the optimization problem of this model is now identical to that of model 1 of the previous section, the optimum solution m and n is given by the same formulae; only V_1^* is substituted for V_1 .

Model 3 - Simple Random Subsampling; m is large relative to M and the estimator of u is of a ratio type. Assume that an auxiliary variable g_k exists for each biomass sample k and that the corresponding total value G of all M biomass sample values g_k is known without error. If the correlation between z_k and g_k is sufficiently strong, an efficient estimator of u is the well known ratio estimator

$$u_r = Gr$$

where, for Σ taken over the sample units $k = 1, 2, \dots, m$,

$$r = \Sigma z_k / \Sigma g_k = \bar{z} / \bar{g}$$

= estimator of the ratio $\rho (= u/G)$ of the true value u of the variable of interest to the true value G of the auxiliary variable.

The estimator u_r is biased, but the bias is expected to be small when m is sufficiently large. The size of the bias can be approximated by the formula

$$B = (M-m) (\rho \sigma_{gg} - \sigma_{zg}) / m \mu_g$$

where $\mu_g = G/M$ is the expected value of g_k , σ_{gg} is the variance of g_k and σ_{zg} is the covariance of z_k and g_k . The variance of u_r as an estimator of the true value u of the given tree is approximately equal to

$$\sigma_{u_r u_r} = \left(\frac{M-m}{M} \right) (V_2^* / m)$$

where, for Σ taken over all tree values $k = 1, 2, \dots, M$,

$$V_2^* = M^2 \Sigma (z_k - \rho g_k)^2 / M = M^2 (\sigma_{zz} - 2\rho \sigma_{zg} + \rho^2 \sigma_{gg})$$

By using statistics g , r , S_{zz} , S_{zg} , and S_{gg} calculated from the m pairs of values z_k and g_k one can define similarly estimators of the bias and variance of u_r .

To better see the meaning of u_r , consider the problem of the estimation of the bole oven-dry weight y of a given tree. We are interested in the regression of y on diameter d and empirical evidence suggests that the conditional variance of y given d is proportional to d^4 . This implies that we should work with the transformed variable $u = y/d^2$, the bole biomass (oven-dry weight) per square inch of diameter at breast height. Let the biomass samples be disks cut from the tree bole, M be the total number of disks for the given tree and m be the number of sample disks selected at random and without replacement. It is known that the green and oven-dry weights are highly correlated and that G = total green weight of the entire bole divided by d^2 is easy to determine. If we define now the disk values z_k and g_k by the formulae

$$z_k = (\text{oven-dry weight of disk } k)/d^2$$

and

$$g_k = (\text{green weight of disk } k)/d^2$$

we have the following statistics and estimators

$$r = \bar{z}/\bar{g} = \text{sample ratio of oven-dry/green weight} \\ = \text{estimator of the ratio } \rho (=u/G) \text{ of the unknown total oven-dry weight } u \text{ to the known green weight of the tree bole, and}$$

$$u_r = Gr = \text{estimator of } u, \text{ the total oven-dry weight per square inch of the given tree bole}$$

$$y_r = d^2 u_r = \text{estimator of } y, \text{ the total oven-dry weight of the given tree bole}$$

We can write now the conditional variance V of u_r given $[x]$ as

$$V = V_1 + \left(\frac{M-m}{M} \right) (V_2^{**}/m) = V_1^{**} + V_2^{**}/m$$

where

$$V_1^{**} = V_1 - V_2^{**}/M = V_1 - M(\sigma_{zz} - 2\rho\sigma_{zg} + \rho^2\sigma_{gg})$$

Consequently, substitution of V_1^{**} and V_2^{**} for V_1 and V_2 of the formulae of the previous section 1, yields the optimum values m and n for the present model 3. These formulae may take a different form when $M = \infty$, the case of infinitely many, overlapping biomass samples. Then, V_1^{**} becomes equal to V_1 (since $V_2^{**}/M = 0$) and the definition of the variances and covariance of z_k and g_k must change accordingly (in terms of expected values).

Model 4 - Stratified Random Subsampling. We shall assume that (i) the tree is divided into L large sections known as strata, (ii) each stratum $h = 1, 2, \dots, L$ is further subdivided into m_h biomass samples, (iii) $m_h \geq 2$ biomass samples are selected from each stratum h by simple random sampling without replacement, and (iv) the samples from various strata are statistically independent. As before, we define y as the biomass of the tree (or some of its components) and u as $(y/\sqrt{w})$, the transformed value of y . In addition, we shall also define the following stratum values

$$y_h = \text{biomass of stratum } h$$

$$u_h = y_h/\sqrt{w} = \text{transformed biomass value of stratum } h$$

$$z_{hk} = (\text{biomass of sample } k \text{ within stratum } h)/\sqrt{w}$$

$$\mu_{zh} = \text{mean of the } M_h \text{ values } z_{hk} \text{ within stratum } h$$

$$\bar{z}_h = \text{mean of the } m_h \text{ sample values } z_{hk} \text{ within stratum } h$$

$$\sigma_{zzh} = \text{variance of the } M_h \text{ values } z_{hk} \text{ within stratum } h$$

$$S_{zzh} = \text{variance of the } m_h \text{ sample values } z_{hk} \text{ within stratum } h$$

$$\hat{u}_h = M_h \bar{z}_h = \text{estimator of } u_h$$

$$\sigma_{\hat{u}\hat{u}h} \approx M_h(M_h - m_h)(\sigma_{zzh}/m_h) \\ = (M_h - m_h)(V_{2h}/m_h)/M_h = \text{variance of } \hat{u}_h$$

$$\text{where } V_{2h} = M_h^2 \sigma_{zzh}$$

$$S_{\hat{u}\hat{u}h} = M_h(M_h - m_h)(S_{zzh}/m_h) = \text{estimator of the variance of } \hat{u}_h$$

Using these values, we can define

$$u_s = \hat{u}_1 + \hat{u}_2 + \dots + \hat{u}_L = \Sigma \hat{u}_h \\ = \text{stratified sampling estimator of the true value } u \text{ of the given tree}$$

and

$$V_s = \sigma_{\hat{u}\hat{u}s} = \Sigma M_h(M_h - m_h)(\sigma_{zzh}/m_h) \\ = \Sigma (M_h - m_h)(V_{2h}/m_h)/M_h = \text{variance of } u_s$$

If we define

$$V_1^s = V_1 - \Sigma (V_{2h}/m_h)$$

we can write the conditional variance of u_r given $[x]$ as

$$V = V_1^s + \Sigma (V_{2h}/m_h)$$

If we further define

$$c_h = \text{average cost of measuring a biomass sample from stratum } h = 1, 2, \dots, L$$

the optimization problem can be expressed in the following two forms

Optimization Problem 1. Find values $n, m_1, m_2, \dots, m_L$ that minimize the cost (objective) function

$$C = nc_t + \Sigma nm_h c_h$$

subject to the side (variance) condition that

$$(V_1^s + \Sigma (V_{2h}/m_h))/n = K_0$$

where all parameters c_t, c_h, V_1^s, V_{2h} , and K_0 are positive values and the optimum solution $n, m_1, m_2, \dots, m_L$ are positive integers (greater than two), and

Optimization Problem 2. Find values $n, m_1, m_2, \dots, m_L$ that minimize the variance (objective) function

$$K = (V_1^s + \Sigma (V_{2h}/m_h))/n$$

subject to the side (cost) condition that

$$nc_t + \Sigma nm_h c_h = C_0$$

where c_t, c_h, V_1^s, V_{2h} , and C_0 are positive values and $n, m_1, m_2, \dots, m_L$ are positive integers (greater than two).

To solve these problems, we shall proceed as follows. It is known, see Cochran (1977) among

others, that for a given, overall sample size $m = (m_1 + m_2 + \dots + m_L)$, the optimum allocation of the sample size m_h within stratum h is proportional to the values $\sqrt{V_{2h}/c_h}$. Note that the effect of the stratum size M_h that is shown explicitly in Cochran's procedure is included implicitly in the definition of V_{2h} , since $V_{2h} = M_h^2 \sigma_{zzh}$. This implies that

$$(1) m_h = m \sqrt{V_{2h}/c_h} / \sum \sqrt{V_{2h}/c_h} \\ = m M_h \sqrt{\sigma_{zzh}/c_h} / \sum M_h \sqrt{\sigma_{zzh}/c_h} \\ (2) \sum m_h c_h = m \sum c_h \sqrt{V_{2h}/c_h} / \sum \sqrt{V_{2h}/c_h} \\ = m \sum \sqrt{V_{2h} c_h} / \sum \sqrt{V_{2h}/c_h} \\ = m \sum M_h \sqrt{\sigma_{zzh} c_h} / \sum M_h \sqrt{\sigma_{zzh}/c_h} \\ = m c_s$$

where $c_s = \sum m_h c_h / m$ = average cost of measuring a biomass sample from the overall population of biomass samples when the allocation is optimum, and

$$(3) \sum (V_{2h}/m_h) = (\sum \sqrt{V_{2h}/c_h}) (\sum \sqrt{V_{2h}/c_h}) / m \\ = (\sum \sqrt{V_{2h}/c_h}) (\sum \sqrt{V_{2h} c_h}) / m \\ = (\sum M_h \sqrt{\sigma_{zzh}/c_h}) (\sum M_h \sqrt{\sigma_{zzh} c_h}) / m \\ = V_2^s / m$$

where

$$V_2^s = (\sum \sqrt{V_{2h}/c_h}) (\sum \sqrt{V_{2h} c_h}) \\ = (\sum M_h \sqrt{\sigma_{zzh}/c_h}) (\sum M_h \sqrt{\sigma_{zzh} c_h})$$

Because the cost and variance functions can now be written as

$C = n c_t + m m c_s$
and
 $K = (V_1^s + V_2^s/m)/n$

respectively, the solution to the two optimization problems of this section is the same as the solution of the two optimization problems of the previous section. Of course, V_1^s and V_2^s are substituted for V_1 and V_2 respectively and the definition of c_s is slightly different.

Remarks

(1) Most of the remarks made in the previous section for model 1 apply here as well. More specifically, (i) all m_h and n values are obtained by the rounding-off (to the nearest or nearest above) integer of the optimum solution values m_h and n obtained by the calculus optimization techniques, (ii) the optimum solution makes sense from an intuitive point of view; m_h increases with the increase of V_{2h} and/or decrease of c_h , and (iii) for practical considerations we have to assume that the variation within trees is the same from tree to tree.

(2) It is extremely difficult, if at all possible, to have good estimates of stratum variances σ_{zzh} , V_{2h} or costs c_h , c_s . Therefore, one can make assumptions that would simplify the formulae. For example, one may assume that $c_h = c_s$ for all h . Then,

$$m_h = m \sqrt{V_{2h}} / \sum \sqrt{V_{2h}}, \text{ and } V_2^s = (\sum \sqrt{V_{2h}})^2$$

Or one may assume that $\sigma_{zzh} = \sigma_{zz}$ in addition to $c_h = c_s$ for all h . Then, for $\sum M_h^2 = M$,

$$m_h = m M_h / M \text{ and } V_2^s = M^2 \sigma_{zz}$$

Finally, if $M_h = M/L$ (equal strata), $\sigma_{zzh} = \sigma_{zz}$ and $c_h = c_s$ for all h , the formulae becomes

$$m_h = m/L \text{ and } V_2^s = M^2 \sigma_{zz} = L^2 V_2$$

since $V_2 = M^2 \sigma_{zzh} = M^2 \sigma_{zz} / L^2$

(3) We have defined V_2^s as the quantity

$$V_2^s = m \sum (V_{2h}/m_h) = \sum (V_{2h}/(m_h/m))$$

where m_h/m were the optimum allocation ratios. Because m_h must be an integer and

$$m_h/m = \sqrt{V_{2h}/c_h} / \sum \sqrt{V_{2h}/c_h}$$

is generally a non-integer, in real life V_2^s , as given by the formula above, will only be an approximation to the optimum value V_2 . This should not be of any practical concern, however, because m_h is taken as the nearest integer, and with sufficiently large m , the actual allocation ratios m_h/m are not too far from the optimum ratios.

(4) If instead of the optimum allocation ratio m_h/m one uses any arbitrary allocation $w_h = m_h/m$, one can still use our approach. However, in this case we define V_2^s by the formula

$$V_2^s = \sum (V_{2h}/w_h)$$

Once m is determined by our optimization formulae, one can apply the allocation ratios w_h to obtain the sample allocations m_h ; that is

$$m_h = (\text{nearest integer } \geq 2) \text{ to } m w_h$$

(5) When for all $h = 1, 2, \dots, L$, the stratum size M_h is large with respect to subsample size m_h , one can simplify the model by making $(M_h - m_h)/M_h = 1$ and, thus, by making $V_1^s = V_1$. The same simplification occurs when the biomass samples are overlapping and M_h is infinite in size. But then, one must use the relative size A_h of stratum h to define $V_{2h} = A_h^2 \sigma_{zz}$; the factors A_h allow one to go from the "per biomass" sample values z to the "per stratum" totals u_h , that is from μ_{zh} to $u_h = A_h \mu_{zh}$.

Model 5 - Stratified Random Subsampling with Ratio Estimators Within Strata.

With this subsampling method, the procedure to select the biomass samples is the same as in Model 4 above but (i) the biomass samples are measured for the auxiliary variables g_{hk} in addition to the variables of interest z_{hk} , (ii) the stratum totals G_h of g_{hk} are known for each h , and (iii) ratio estimators of the form $u_{rh} = G_h r_h$ (defined in Model 3) are used for u_h (instead of $\hat{u}_h$). Using the procedure and notation of Model 3, we can write

$$r_h = \sum z_{hk} / \sum g_{hk} = \bar{z}_h / \bar{g}_h \\ = \text{estimator of the ratio } \rho_h (= u_h/G_h) \text{ of} \\ \text{the true value } u_h \text{ of the variable of} \\ \text{interest to the true value } G_h \text{ of the} \\ \text{auxiliary variable for stratum } h$$

$$B = (M_h - m_h) (\rho_h \sigma_{ggh} - \sigma_{zgh}) / m_h \mu_{gh} \\ = \text{bias of ratio estimator } u_{rh} \text{ of } u_h,$$

where $\mu_{gh} = G_h/M_h$ = mean of auxiliary variable g_{hk} ,

σ_{ggh} is the variance of g_{hk} , and σ_{zgh} is the covariance of z_{hk} and g_{hk} within stratum h ; and

$$\sigma_{u_{rh}u_{rh}} = \left(\frac{M_h - m_h}{M_h} \right) (V_{2h}^* / m_h)$$

= variance of u_{rh} as an estimator of u_h

where

$$V_{2h}^* = M_h^2 \Sigma (z_{hk} - \rho g_{hk})^2 / M_h = M_h^2 (\sigma_{zzh} - 2\rho \sigma_{zgh} + \rho^2 \sigma_{ggh})$$

Define now

$$u_s = u_{r1} + u_{r2} + \dots + u_{rL} = \Sigma u_{rh}$$

= estimator of the true biomass value u of the given tree

and

$$V_s = \Sigma \sigma_{u_{rh}u_{rh}} = \Sigma (M_h - m_h) (V_{2h}^* / m_h) / M_h$$

= variance of u_s

If we define

$$V_1^* = V_1 - \Sigma (V_{2h}^* / m_h)$$

we can write the conditional variance of u_{rh} given $[x]$ as

$$V = V_1^{s*} + \Sigma (V_{2h}^* / m_h)$$

The optimization problems for this model are the same as for the previous Model 4, with V_1^{s*} and V_{2h}^* substituted for V_1 and V_{2h} respectively. Consequently, the optimum sample sizes m_h and n can now be determined by the procedure outlined for Model 4. We have

$$m_h = m \sqrt{V_{2h}^* / c_h} / \sqrt{V_{2h}^* / c_h} \quad \text{for } h = 1, 2, \dots, L$$

$$\Sigma m_h c_h = mc_s$$

$$\text{where } c_s = \Sigma \sqrt{V_{2h}^* / c_h} / \sqrt{V_{2h}^* / c_h}$$

and

$$\Sigma (V_{2h}^* / m_h) = (\Sigma \sqrt{V_{2h}^* / c_h}) (\Sigma \sqrt{V_{2h}^* / c_h}) / m$$

= V_2^{s*} / m

Consequently, the cost and variance function can be written as

$$C = nc_t + mnc_s$$

and

$$K = (V_1^{s*} + V_2^{s*} / m)$$

respectively, and the optimum values m and n are given by the formulae of Model 1.

Summary Comments

We have considered the problem of optimizing the size of the subsample used to estimate the biomass of sample trees. We have assumed that the n sample trees were selected by simple random sampling, that each sample tree was subdivided into M biomass samples (small sections of the tree) and that m biomass samples were selected by simple random sampling without replacement and measured for the biomass variable of interest z and possibly for a second variable, denoted here as the auxiliary variable g , highly correlated with z . The selection of the n sample trees and the n subsamples (within the n trees) were all assumed to be statistically independent of each other.

To solve this problem, we have constructed an optimization model of the mathematical programming form, having an objective function to minimize and side conditions to satisfy. The decision variables are the sample and subsample size n and m , the parameters of the model are (i) the variance components V_1 (due to the variation of the true biomass about its regression function), and V_2 (due to the variation of the biomass sample values within the sample trees), (ii) the average sampling costs c_t per sample tree and c_s per biomass sample, and (iii) either a desired precision K_0 of some biomass estimators or an allowable cost of sampling C_0 . We have considered the two essentially equivalent problems of (i) minimizing the sampling costs C subject to the side condition that the precision of the parameter of interest is equal to K_0 , or (ii) minimizing the variance of the estimator of interest subject to the side condition that the sampling costs are equal to C_0 . The model had the additional side condition that the optimum sample sizes n and m must be positive integers.

After formulating the mathematical models, we have solved the models by calculus techniques of optimization. The solution shows that the optimum size m of the subsample does not depend on the number n of sample trees; it depends only on the variance components V_1 and V_2 as well as the cost parameters c_t and c_s . Once the subsample size m is determined, then the number n of sample trees is a function of the total variance function V (of the form $V_1 + V_2/m$) and either the desired precision $V/n = K_0$ or the allowable costs of sampling $n(c_t + mc_s) = C_0$.

Of course, the subsample size m depends also on the design used to subsample the tree biomass. We have started with the most simple case where the subsampling is done by simple random sampling with replacement or simple random sampling without replacement where the effect of the finite population correction factor can be ignored. We have then extended this model to include the effect of the finite population correction factor (when it cannot be ignored) and consider the case where the estimator of the tree biomass that uses the subsample average $\bar{z}$ is replaced by a ratio type estimator. We have further extended the model by replacing the simple random by the stratified random sampling method of selecting the m biomass samples. We have also considered both estimators of the tree biomass; the one based on the means of various strata and the one based on ratio type estimators of the stratum totals. In all these extensions we have shown how to change the definition of the variance components V_1 and V_2 so that (i) the new components are consistent with sampling design and estimator used and (ii) the mathematical model to optimize does not change and the optimum solution of m and n remains the same function of the (new) components V_1 and V_2 .

The application of this approach to real life problems may present several difficulties or practical limitations. It is first necessary to obtain sufficiently good estimates of the average cost factors c_t and c_s as well as estimates of the variance components V_1 and V_2 . And these may be difficult to obtain. Some sensitivity analysis of the

model may be needed to determine how much variation in the estimates of c_t , c_s , V_1 , and V_2 can be tolerated without critically affecting the optimum solution. Furthermore, to keep the mathematical model sufficiently simple, we had also to assume that the variance component V_2 is the same (or does not change much) from tree to tree. Because (i) V_2 refers to the variance of the ratio $u = y/\sqrt{w}$ of the total biomass y of some tree component to some known weighting factor $\sqrt{w}$ (a function of tree size and possibly species), and (ii) the variance of u is expected to be approximately homogeneous, it may imply that the assumption that V_2 does not vary much from tree to tree is not too critical.

For example, let y be the oven-dry biomass of the main tree bole and assume that the conditional variance of y given diameter d and height h is approximately proportional to $d^4 h^2$. Then, the variable $u = y/d^2 h$, which has an approximately homogeneous variance, has a special meaning that may imply that the variation of V_2 from tree to tree is expected to be small. This can be explained as follows. It is known that the bole volume is approximately equal to some factor k multiplied by the quantity $d^2 h$. Note that $\pi d^2/4h$ represents the volume of a cylinder of diameter d and length h and $k d^2 h$ represents the volume of the tree bole as a fraction of the corresponding cylinder. Consequently, dividing y by $d^2 h$ is nothing but expressing the biomass value y on a "per volume unit" basis. It is not necessary for the volume unit to be exactly a cubic foot; any unit would suffice. As (i) z is the M -th part of the value u of the main bole, and (ii) z , as a relative value represents "weight per volume unit," varies little from tree to tree, it seems reasonable to expect that the variation of σ_{zz} or $V_2 = M^2 \sigma_{zz}$, if any, would be small.

The optimization model requires that M be the same from tree to tree. If m is small relative to M and the effect of the finite population correction factor can be ignored, the formulae of the optimum sample sizes m and n are not a function of M . This seems to imply that as long as M remains large with respect to m , one may use different values of M for different trees. This may affect slightly the variance term σ_{zz} and V_2 but it is not expected to affect it critically.

The approach shown here can be extended to other methods of subsampling. It suffices to identify the form that V_1 and V_2 will take with these methods. The modification of the sampling technique to select the n sample trees may, however, change critically the optimization model. Instead of one sample size n , the model may have to be written as a function of several sample sizes $n_1, n_2, \dots$. For example, if the sample trees are selected by cluster sampling, the sample sizes may be n_1 = number of clusters and n_2 = number of trees per cluster, assuming that n_2 is the same from cluster to cluster.

Acknowledgments

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

- Cochran, W. G. Sampling Techniques, third edition. John Wiley and Sons, Inc., New York, NY. 1977.
- Cunia, T. On the error of biomass estimates in forest inventories: Part 1: Its major component. Faculty of Forestry Miscellaneous Publication Number 8, SUNY College of Environmental Science and Forestry, Syracuse, NY. 1986a.
- Cunia, T. On the error of biomass estimates in forest inventories: Part 2: The error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9, SUNY College of Environmental Science and Forestry, Syracuse, NY. 1986b.
- Valentine, H.T.; Tritton, L. M.; Furnival, G. M. Subsampling trees for biomass, volume or mineral content. Forest Science 30: 673-681. 1984.

ESTIMATING SAMPLE TREE BIOMASS BY SUBSAMPLING:

SOME EMPIRICAL RESULTS

R. D. Briggs, T. Cunia, E. H. White and H. W.

Yawney

Post-doctoral Research Associate, Professor of Statistics and Operations Research, Professor at SUNY College of Environmental Science and Forestry, Syracuse, NY, and Research Scientist, USDA-Forest Service, Northeastern Forest Experiment Station, Burlington, VT respectively

A procedure used to measure the green weight and estimate, by subsampling, the dry weight of the above ground components of 29 randomly selected sugar maple trees is described. Foliage and branch dry weights were determined by direct measurement. Bole wood and bole bark dry weight was estimated by stratified subsampling and subsequent application of ratio-type estimators. An analysis was made of the variation of subsamples within sample trees and variance components are calculated and listed for use in later studies.

Introduction

Regression functions of tree biomass on diameter outside bark at breast height (dbh) are estimated from samples of trees randomly selected from some forest population of interest. Once these trees are selected, their biomass must be measured, if possible, or estimated by subsampling otherwise. If the biomass is expressed as green weight, the tree component can be measured directly as soon as it is separated from the tree. The only possible error may be due to faulty measurement instruments or methods. If the biomass is expressed as dry weight, direct measurement may be too expensive and time consuming for large tree components. The only practical alternative is subsampling; small biomass samples are selected from the tree component by some random procedure. Green and oven-dry weights of these samples are determined in the laboratory and the results extrapolated from biomass samples to the entire tree component.

Note that we define here the "measurement" of biomass as the process of direct determination of the biomass of the entire tree component of interest and the "estimation" of biomass as the process of determination of the biomass by subsampling. The fresh weight of a component is measured if the entire component is weighed or estimated if parts of the component are weighed and the weights of the parts are extrapolated, by some computational procedure, to the entire component. We also define here the measurement error as the difference between the actual value as obtained by the measurement or estimation process and the true value of the parameter of

interest. We shall assume that this error is negligibly small when the biomass is measured but is sufficiently large when the biomass is estimated. The measurement error can be evaluated and expressed in statistical terms if the subsampling procedure is statistical.

The objective of the present paper is to describe the procedure used to measure and estimate the biomass of the aboveground components of 29 sugar maple (*Acer saccharum* Marsh.) trees selected at random from a 21 ha sugar maple stand on the Green Mountain National Forest in the town of Bristol, Addison County, Vermont. These trees were part of a large study designed to evaluate the effects of whole-tree and tree-length harvesting on future site productivity of northern hardwoods. Additional details of this study are provided by Briggs (1985). More specifically, we shall be concerned here with the measurement error due to the process of determining, by subsampling, the biomass of the sample trees. Different people, applying the same or different subsampling procedures will generally arrive at different tree biomass values and thus, arrive at different regression functions.

Subsampling Procedure for Estimating Tree Biomass

The 29 sugar maple trees were all measured first for their diameter at breast height (dbh), total height (h), and merchantable height to a 10 cm minimum top diameter. Each tree was then felled down and ten plastic sheets were placed on the ground surrounding the tree. Beginning at the base of the crown and working towards the top, the tree branches with their leaves attached were removed in groups of ten, so that each branch in a given group had approximately the same amount of branchwood and foliage (as estimated ocularly) and came from a similar height in the crown. One branch from each group was put in a different pile so that for each tree, ten piles were obtained; each pile having a similar distribution of branches and foliage with respect to weight and point of origin from the crown.

For each of the ten piles, all of the foliage was picked from the branches, placed in paper bags and carried to the truck. The branches from each pile were tied together in canvas tarpaulins and carried to the landing. In order to facilitate transportation and oven-drying, each pile of branches was processed separately through a small chipper operated from the PTO (power take-off) shaft of a farm tractor. The chips were collected and bagged separately by pile and transported to the laboratory.

The bole of each sample tree was divided into three sections of equal length. For each section, three integers were randomly selected from 1 to 100. Each of these numbers was multiplied (as a percent) by the section length to obtain the location of a sample disk for the determination of the fresh and dry weight. For example, if the random number was 32 and the section length was 5.0 m, then a disk would be

located at $(.32) \times (5.0) = 1.50$ m from the base of that section. Each disk location was marked with flagging.

Each of the three bole sections was cut into logs of various lengths and weighed on a 90 kg capacity field scale. Disks, approximately 5 cm in width, were removed from the bole (at the marked locations) after each section was completely weighed, placed in double plastic bags and carried to the truck. At the end of each day, the sample disks were weighed in the field office and transferred to paper bags.

All sample material (foliage, branches, disks) was transported to the Forest Service Laboratory in Burlington, Vermont and placed in a greenhouse. The greenhouse was used as a holding area which allowed for partial drying, preventing losses of dry weight due to microbial decomposition. At the end of the field season, all sample material was transported to the SUNY College of Environmental Science and Forestry, Syracuse, N.Y. Foliage, branches and disks were placed in forced air kilns at 65° C until constant weight was obtained. The oven-dry weight was determined

individually for each pile of branches and foliage, as well as for each disk of each individual sample tree. The bark was removed from each disk, dried at 65° C and its weight was recorded.

The measured tree data including dbh, total height and height to a 10 cm minimum top diameter, total bole length and the green weight (wood and bark) for each of the three bole sections are shown in (Table 1). The green and dry weight data for wood and bark of each of the nine sample disks per tree can be found in Briggs et al. (1986). Finally, the mean oven-dry weights for the ten piles of foliage and branches, along with their variances, are shown in Table 2.

Estimation of Bole Biomass

Determination of the green weight of the tree bole is made by direct measurement of its weight. If the loss of humidity between tree felling and bole weighing is negligibly small and the weighing scale is accurate, the measurement error of the bole biomass, expressed as green

Table 1. Morphological data for the 29 sample sugar maple trees.

Tree	Dbh cm	Total Height m	Height to 10 cm Top m	Bole Length m	Green Weight of Bole (wood + bark)			
					Bot	Mid	Top	Total
					-----kg-----			
1	25.1	15.7	15.4	14.8	196.8	142.1	76.0	414.9
2	43.4	19.9	14.0	13.9	595.5	493.9	179.3	1268.7
3	33.0	20.7	15.7	15.5	374.4	316.8	174.3	865.5
4	41.9	20.4	15.2	14.7	544.8	415.9	223.5	1184.2
5	17.6	19.5	11.9	11.5	90.1	77.3	50.1	217.5
6	16.7	19.8	12.4	12.1	75.4	52.5	33.0	160.9
7	38.1	23.5	18.2	17.0	526.1	434.0	204.3	1164.4
8	10.2	16.4	4.4	4.1	10.9	9.1	7.9	27.9
9	32.5	23.2	19.9	19.6	475.5	345.2	131.7	952.4
10	38.6	22.6	18.1	17.6	550.3	408.4	164.6	1123.3
11	41.2	23.4	17.6	17.3	603.0	428.8	239.2	1271.0
12	25.4	21.7	15.2	14.9	196.2	133.7	77.9	407.8
13	32.5	22.0	17.4	17.0	426.8	275.0	142.8	844.6
14	9.1	13.0	2.2	2.2	5.0	5.4	5.4	15.8
15	32.5	24.0	17.8	17.4	417.1	340.4	205.9	963.4
16	25.9	22.0	16.3	16.0	238.0	162.7	92.4	493.1
17	9.1	12.4	1.8	1.5	3.8	3.1	2.8	9.7
18	17.2	18.5	10.7	10.4	74.4	47.2	29.9	151.5
19	39.4	21.8	16.4	16.1	597.5	418.4	235.4	1251.3
20	26.2	22.0	15.6	15.3	233.1	164.5	106.4	504.0
21	37.3	23.4	17.4	17.0	508.3	393.9	189.9	1092.1
22	43.9	22.2	20.8	20.4	812.5	567.9	185.4	1565.8
23	38.6	23.5	18.5	18.0	631.6	493.3	244.5	1369.4
24	9.9	13.7	3.5	3.1	11.3	4.1	6.8	22.2
25	26.4	22.8	16.5	16.1	259.1	175.5	106.3	540.9
26	33.0	22.2	17.6	17.3	428.4	317.3	185.4	931.1
27	42.7	25.2	18.4	18.0	768.6	656.6	420.3	1845.5
28	38.9	23.5	18.4	17.9	589.7	435.4	191.5	1216.6
29	42.2	25.4	19.9	19.5	657.0	487.8	314.4	1459.2

Table 2. Total and average dry weights for the ten piles of foliage and branches for each of the 29 sample sugar maple trees.

Tree	Foliage			Branches		
	Total	Average	Variance Between Piles	Total	Average	Variance Between Piles
	-----kg-----		kg ²	-----kg-----		kg ²
1	7.48	0.748	0.014856	81.56	8.156	3.875084
2	24.10	2.410	0.076900	363.98	36.398	16.106596
3	16.09	1.609	0.113369	126.88	12.688	17.074796
4	16.03	1.603	0.149081	204.44	20.444	6.623464
5	3.48	0.348	0.009996	45.09	4.509	1.248909
6	2.89	0.289	0.006409	28.47	2.847	1.610921
7	12.68	1.268	0.044956	208.28	20.828	13.249076
8	0.61	0.061	0.000369	21.77	2.177	0.282821
9	11.07	1.107	0.053941	130.81	13.081	2.884529
10	13.55	1.355	0.090625	176.90	17.690	4.592060
11	14.23	1.423	0.264621	178.37	17.837	17.018581
12	5.83	0.583	0.020241	64.06	6.406	3.130104
13	8.07	0.807	0.033021	110.69	11.069	4.050209
14	0.68	0.068	0.001416	21.30	2.130	0.880420
15	8.56	0.856	0.014764	87.79	8.779	2.467629
16	4.74	0.474	0.018304	67.50	6.750	3.688480
17	0.56	0.056	0.000384	20.30	2.030	0.458400
18	2.41	0.241	0.002629	26.94	2.694	0.246784
19	25.81	2.581	0.228229	335.42	33.542	32.040436
20	6.47	0.647	0.016781	70.14	7.014	2.392004
21	3.81	0.381	0.006669	121.29	12.129	6.776149
22	23.11	2.311	0.056009	295.50	29.550	5.153040
23	11.96	1.196	0.037364	175.85	17.585	21.395305
24	0.73	0.073	0.000641	19.26	1.926	0.793304
25	6.53	0.653	0.023201	87.24	8.724	5.915184
26	14.06	1.406	0.037444	153.75	15.375	1.502325
27	25.47	2.547	0.300041	482.82	48.282	47.007136
28	8.65	0.865	0.040385	163.33	16.333	16.919021
29	8.85	0.885	0.018745	185.29	18.529	15.937229

weight, is equal to zero. On the other hand, the ovendry weight of the bole is determined by sub-sampling. As the three bole sections can be considered as strata, and three disks are selected at random from each section, the method of disk selection is stratified random sampling. Because the green weight of the entire bole, individual sections and disks are known and the ovendry weight of the sample disks is measured, one can estimate ovendry weight of the bole by a stratified ratio estimator.

Let us show in detail the estimation procedure for the ovendry weight of the bole (wood plus bark). The estimation of the ovendry weight of wood alone or bark alone would follow the same general lines. It is then sufficient to show only which variables are substituted into the corresponding formulae. The following notation and definitions of estimators are being used:

G = total green weight of the entire tree bole

G_h = green weight of section h (1 for bottom, 2 for middle and 3 for top section)

g_{hk} = green weight of wood and bark of disk $k = 1, 2, 3$ in stratum h

d_{hk} = ovendry weight of wood and bark of disk $k = 1, 2, 3$ in stratum h

$m_h = 3$ = number of sample disks in each section $h = 1, 2, 3$

$\bar{g}_h = \Sigma g_{hk} / n_h = (g_{h1} + g_{h2} + g_{h3}) / 3$

= average green weight per sample disk within stratum $h = 1, 2, 3$

$$\bar{d}_h = \Sigma d_{hk} / n_h = (d_{h1} + d_{h2} + d_{h3}) / 3$$

= average oven-dry weight per sample disk within stratum $h = 1, 2, 3$

$$M_h = G_h / \bar{g}_h = \text{conceptual number (not necessarily integer) of disks of average weight } \bar{g}_h \text{ that can be made from section } h = \text{size of section } h$$

$$(M_h - m_h) / M_h = \text{finite population correction factor of section } h$$

$$S_{dd,h} = \Sigma (d_{hk} - \bar{d}_h)^2 / (m_h - 1) = \text{sample variance of the } m_h \text{ disk values } d_{hk} \text{ within section } h$$

$$S_{gg,h} = \Sigma (g_{hk} - \bar{g}_h)^2 / (m_h - 1) = \text{sample variance of the } m_h \text{ disk values } g_{hk} \text{ within section } h$$

$$S_{dg,h} = \Sigma (d_{hk} - \bar{d}_h)(g_{hk} - \bar{g}_h) / (m_h - 1) = \text{sample covariance of the } m_h \text{ pairs of disk values } d_{hk} \text{ and } g_{hk} \text{ within section } h$$

$$r_h = \bar{d}_h / \bar{g}_h = \text{estimator of the ratio of the oven-dry weight to the green weight of section } h$$

$$D_h = G_h r_h = M_h \bar{d}_h = \text{ratio estimator of the oven-dry weight of section } h$$

$$B_h = (M_h - m_h)(r_h S_{gg,h} - S_{dg,h} / m_h \bar{g}_h) = \text{estimator of the bias of } D_h$$

$$S_{DD,h} = M_h (M_h - m_h) (S_{dd,h} - 2r_h S_{dg,h} + r_h^2 S_{gg,h}) / m_h$$

= estimator of the variance of D_h .

Note that, in the formulae above, the section (stratum) size M_h is a random variable; it is defined as a function of the random variable $\bar{g}_h$. Furthermore the disk material is not selected with equal probability; the material from the smaller end diameter of the bole is sampled with a higher probability. Consequently, our "ratio" estimator is somewhat different from the classical ratio estimators as they are commonly defined in standard textbooks, where M_h is fixed and the sampling is done with equal probability.

Because D_1 , D_2 and D_3 are statistically independent random variables, the oven-dry weight of the entire tree bole and its error can be estimated by the following formulae:

$$D = \Sigma D_h = D_1 + D_2 + D_3 = \text{stratified ratio estimator of the oven-dry weight of the wood and bark of the tree bole}$$

$$B = \Sigma B_h = B_1 + B_2 + B_3 = \text{estimator of the bias of } D$$

$$S_{DD} = \Sigma S_{DD,h} = \text{estimator of the variance of } D$$

$$\sqrt{S_{DD}} = \sqrt{\Sigma S_{DD,h}} = \text{estimator of the standard error of } D, \text{ and}$$

$$D \pm 2\sqrt{S_{DD}} = \text{approximate 95 percent confidence limits of the oven-dry weight of the tree bole}$$

If we define now:

$$d_{hk,w} = \text{oven-dry weight of the wood of disk } k \text{ of section } h, \text{ and}$$

$$d_{hk,b} = \text{oven-dry weight of the bark of disk } k \text{ of section } h, \text{ and}$$

if these values are substituted for the variable d_{hk} of the formulae above, one can define the estimators D_w and D_b , the stratified ratio estimators of the oven-dry weight of wood and bark, respectively, of the tree bole, as well as the corresponding estimators of their biases, variances and confidence limits.

In order to conserve space and still provide the necessary details required to demonstrate the estimation procedure for sample tree biomass and its associated error, intermediate statistics are provided for the bole component of only 3 of the 29 sample trees (1, 14, 29). The statistics $\bar{g}_h$, $\bar{d}_h$, $S_{dd,h}$, $S_{dg,h}$, $S_{gg,h}$, r_h , D_h , B_h and $S_{DD,h}$ of the oven-dry weight of the wood and bark for the bottom, middle and top sections of the bole for the three selected sample trees are shown in Table 3. The summary statistics D , B , S_{DD} , $\sqrt{S_{DD}}$ and the approximate 95 percent confidence limits for the bole are shown for all 29 trees in Table 4. The summary statistics for the bolewood and bole-bark oven-dry weights are shown in Tables 5 and 6, respectively.

Analysis of Biomass Data and Their Error

This study is concerned with the measurement of the biomass of tree foliage, branches, bole-wood, bole-bark and bole (wood + bark). Only the last tree component was measured for its green weight. We expect a small loss of humidity, if any, between the time of tree felling and bole weighing, as well as a small loss, if any, of wood or branch material due to bole sectioning prior to measurement of green weight. We also expect a sufficiently high accuracy and precision of the field scale used and we do not expect that any blunders were made by the field people when the bole was weighed, (i.e., misreading the scale or error in recording weights). Consequently, we should assume here that the green weight of the main tree bole as shown in Table 1 represents a true value without any measurement error.

Similarly, we shall also assume that the oven-dry weight of foliage and branches as shown in Table 2 has no measurement error. This means that (i) the possible loss of foliage or branch material from the time that the tree was felled in the field to the time of measurement in the laboratory is negligibly small, (ii) the measuring instruments have high precision and accuracy and (iii) no blunders were made when the oven-dry weight was measured and recorded.

Because the foliage and branch material was divided into ten piles and the biomass of each pile was measured and recorded separately by pile, it is possible to make inferences about the

Table 3. The statistics associated with the estimation of the bole oven-dry weight for the three sections of three of the twenty nine sample trees.

Tree	$\bar{d}_h$	$\bar{g}_h$	$S_{dd,h}$	$S_{dg,h}$	$S_{gg,h}$	r_h	D_h	B_h	$S_{DD,h}$
	-----kg-----		-----kg ² -----				-----kg-----		kg ²
Bottom									
1	1.13173	1.79877	0.10058230	0.14796983	0.21774336	0.629	123.82	-0.216	2.2534
14	0.22237	0.35087	0.00035754	0.00033079	0.00044217	0.634	3.17	-0.001	0.0062
29	1.48857	2.45100	0.03355816	0.07364185	0.16421700	0.607	399.02	0.941	110.8305
Middle									
1	0.77747	1.25107	0.00398492	0.00903385	0.02048281	0.621	88.31	0.109	2.7933
14	0.35707	0.57377	0.01694665	0.02708867	0.04334362	0.622	3.36	0.000	0.0003
29	1.35097	2.15000	0.07199929	0.13392440	0.25207900	0.628	306.51	0.849	54.5814
Top									
1	0.47660	0.77373	0.01720093	0.02637653	0.04049172	0.616	46.81	-0.059	0.2181
14	0.22793	0.37467	0.00020024	0.00029754	0.00044804	0.608	3.29	0.000	0.0002
29	0.54330	0.90333	0.18934599	0.32075240	0.54350933	0.601	189.09	0.781	4.9239

Table 4. Summary statistics associated with the calculation of D = estimate of the oven-dry weight for the entire tree bole (wood + bark) for the 29 sample sugar maple trees.

Tree	D	Bias	S_{DD}	$\sqrt{S_{DD}}$	95% Confidence Limits	
					Lower	Upper
	-----kg-----		kg ²		-----kg-----	
1	258.94	-0.166	5.2648	2.29	254.36	263.52
2	782.25	-0.189	38.8910	6.24	769.77	794.73
3	563.30	0.321	599.3665	24.48	514.34	612.26
4	716.41	11.652	1633.0545	40.41	635.59	797.23
5	134.28	0.070	1.3178	1.15	131.98	136.58
6	101.92	0.035	1.6132	1.27	99.38	104.46
7	734.93	-1.154	39.3400	6.27	722.39	747.47
8	17.70	0.000	0.0104	0.10	17.50	17.90
9	599.80	1.032	140.7951	11.87	576.06	623.54
10	727.80	-0.030	83.8443	9.16	709.48	746.12
11	809.70	1.022	48.4170	6.96	795.78	823.62
12	254.41	0.033	8.8941	2.98	248.45	260.37
13	575.41	-1.143	44.5967	6.68	562.05	588.77
14	9.82	-0.001	0.0067	0.08	9.66	9.98
15	600.49	-0.931	338.7310	18.40	563.69	637.29
16	318.11	-0.065	10.5999	3.26	311.59	324.63
17	6.22	-0.001	0.0007	0.03	6.16	6.28
18	90.27	0.003	0.3705	0.61	89.05	91.49
19	777.32	-0.799	59.1263	7.69	761.94	792.70
20	316.66	0.485	30.1318	5.49	305.68	327.64
21	714.69	0.341	192.0900	13.86	686.97	742.41
22	994.01	1.104	185.1251	13.61	966.79	1021.23
23	856.00	1.113	45.3786	6.74	842.52	869.48
24	14.06	-0.050	0.0187	0.14	13.78	14.34
25	339.61	0.372	27.6807	5.26	329.09	350.13
26	591.61	1.260	346.5537	18.62	554.37	628.85
27	1158.86	-1.661	230.1460	15.17	1128.52	1189.20
28	767.45	0.178	140.3759	11.85	743.75	791.15
29	894.62	2.571	170.3358	13.05	868.52	920.72

Table 5. Summary statistics associated with the calculation of D = estimate of the oven-dry weight for the entire tree bolewood for the 29 sample sugar maple trees.

Tree	D	Bias	S _{DD}	$\sqrt{S_{DD}}$	95% Confidence Limits	
					Lower	Upper
	-----kg-----		kg ²		-----kg-----	
1	234.80	-0.253	5.8342	2.42	229.96	239.64
2	710.88	-0.998	47.6966	6.91	697.06	724.70
3	506.01	0.314	591.9527	24.33	457.35	554.67
4	639.79	8.419	1090.6481	33.02	573.75	705.83
5	122.11	0.028	0.5103	0.71	120.69	123.53
6	90.55	0.020	1.1864	1.09	88.37	92.73
7	647.23	-1.378	46.9841	6.85	633.53	660.93
8	16.08	-0.001	0.0007	0.03	16.02	16.14
9	545.02	0.153	84.5232	9.19	526.64	563.40
10	659.37	-0.506	29.4932	5.43	648.51	670.23
11	731.39	0.778	42.1788	6.49	718.41	744.37
12	226.39	-0.161	10.4214	3.23	219.93	232.85
13	505.28	-0.594	32.1162	5.67	493.94	516.62
14	8.33	-0.002	0.0054	0.07	8.19	8.47
15	535.40	-1.491	326.4250	18.07	499.26	571.54
16	288.13	-0.182	11.6678	3.42	281.29	294.97
17	5.59	-0.001	0.0007	0.03	5.53	5.65
18	77.99	-0.030	0.7554	0.87	76.25	79.73
19	690.23	-0.548	21.2706	4.61	681.01	699.45
20	287.61	0.306	27.1924	5.21	277.19	298.03
21	630.19	0.072	167.6352	12.95	604.29	656.09
22	904.14	0.876	182.4555	13.51	877.12	931.16
23	768.70	1.241	54.5529	7.39	753.92	783.48
24	12.36	-0.049	0.0188	0.14	12.08	12.64
25	306.27	0.236	25.2890	5.03	296.21	316.33
26	535.29	0.876	266.2677	16.32	502.65	567.93
27	1040.24	-1.599	252.8985	15.90	1008.44	1072.04
28	690.02	-0.212	141.0389	11.88	666.26	713.78
29	799.75	1.767	131.0079	11.45	776.85	822.65

Table 6. Summary statistics associated with the calculation of D = estimate of the oven-dry weight for the entire tree bolebark for the 29 sample sugar maple trees.

Tree	D	Bias	S _{DD}	$\sqrt{S_{DD}}$	95% Confidence Limits	
					Lower	Upper
	-----kg-----		kg ²		-----kg-----	
1	24.14	0.087	0.2401	0.49	23.16	25.12
2	71.37	0.809	10.2646	3.20	64.97	77.77
3	57.29	0.008	0.7573	0.87	55.55	59.03
4	76.63	3.232	91.8978	9.59	57.45	95.81
5	12.18	0.041	0.2591	0.51	11.16	13.20
6	11.36	0.015	0.2485	0.50	10.36	12.36
7	87.70	0.226	6.4143	2.53	82.64	92.76
8	1.61	0.001	0.0069	0.08	1.45	1.77
9	54.78	0.880	31.8935	5.65	43.48	66.08
10	68.43	0.477	44.3247	6.66	55.11	81.75
11	78.31	0.244	1.5695	1.25	75.81	80.81
12	28.02	0.194	0.8577	0.93	26.16	29.88
13	70.12	-0.549	60.7034	7.79	54.54	85.70
14	1.49	0.001	0.0001	0.01	1.47	1.51
15	65.09	0.560	4.1230	2.03	61.03	69.15
16	29.97	0.116	0.4512	0.67	28.63	31.31
17	0.63	0.000	0.0000	0.00	0.63	0.63
18	12.28	0.033	0.1117	0.33	11.62	12.94
19	87.09	-0.250	18.0980	4.25	78.59	95.59
20	29.06	0.178	0.7317	0.86	27.34	30.78
21	84.50	0.268	1.5386	1.24	82.02	86.98
22	89.87	0.228	10.6607	3.27	83.33	96.41
23	87.31	-0.128	1.2501	1.12	85.07	89.55
24	1.69	0.000	0.0000	0.00	1.69	1.69
25	33.35	0.136	0.9099	0.95	31.45	35.25
26	56.32	0.385	7.1535	2.67	50.98	61.66
27	118.63	-0.061	11.8567	3.44	111.75	125.51
28	77.43	0.391	4.8309	2.20	73.03	81.83
29	94.87	0.804	7.4899	2.74	89.39	100.35

measurement error when a few rather than all piles are being measured. If $m < 10$ piles are selected by simple random sampling without replacement, the biomass of the tree foliage or branches can be estimated by $Y = 10 \bar{y}$, where Y is the estimate of the tree biomass and $\bar{y}$ is the average of the biomass of the m sample piles. The variance of Y can be estimated by the formula

$S_{yy} = (10-m) S_{yy}/10 m$, where S_{yy} is the variance of the estimator Y and S_{yy} is the variance of the biomass values y of the 10 individual piles. Let us analyze more closely the behavior of S_{yy} for both foliage and branch biomass. Because the piles are formed more or less arbitrarily, and it is natural to expect variation in the pile biomass from tree to tree, the values S_{yy} can be viewed as random variables.

Casual examination of the values S_{yy} for the foliage and branch biomass of the 29 sample trees suggests that S_{yy} and tree size are correlated; as the tree increases in size, so does its variance (Tables 1 and 2). This is not surprising since it is empirically known that large things tend to vary more than small things. To better see this relationship, foliage variance S_{yy} was plotted against dbh (Figure 1) and against foliage biomass Y (Figure 2). In both cases the variance S_{yy} seems to increase exponentially with tree size.

It was also felt that the variance between pile biomass may be proportional to the fourth power of tree dbh. As this hypothesis is identical to the hypothesis that the standard deviation is proportional to squared dbh, Figure 3 shows the graph of standard deviation against the squared dbh value. As the reader can verify, this hypothesis seems to be approximately true; a straight line passing through the origin of the two axes seems to be a sufficiently good approxi-

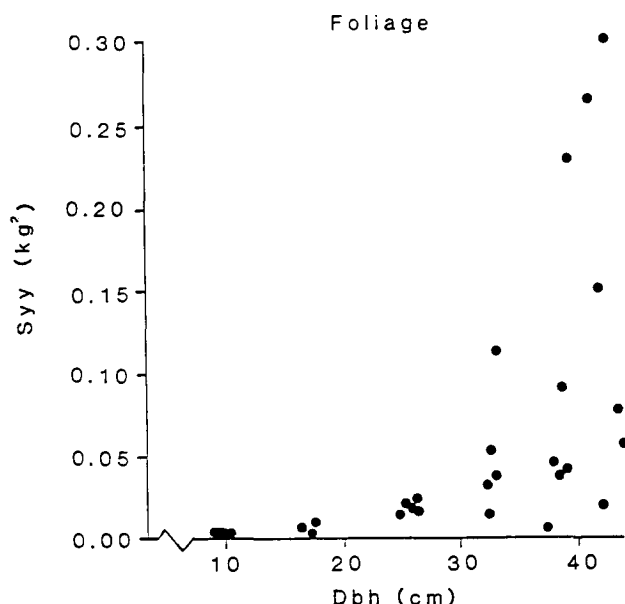


Figure 1. The variance of the foliage biomass (ovendry weight) of the ten piles plotted against the tree diameter.

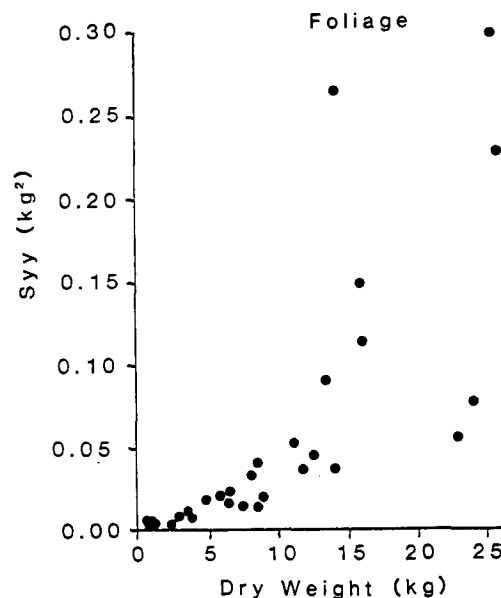


Figure 2. The variance of foliage biomass (ovendry weight) of the ten piles plotted against the total ovendry weight of the tree foliage.

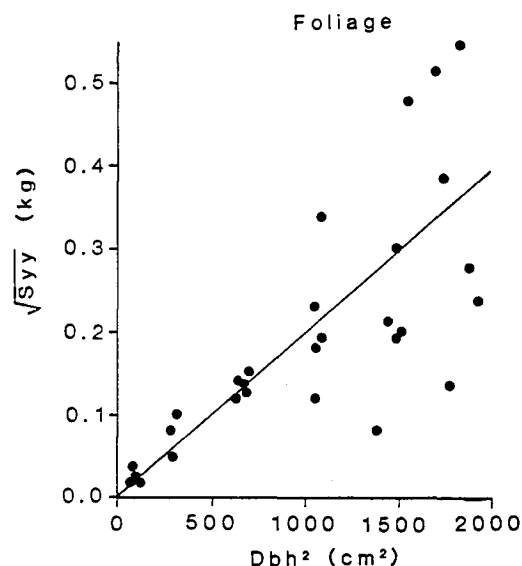


Figure 3. Standard deviation of the foliage biomass (ovendry weight) of the ten piles plotted against the squared value of tree diameter.

mation of this relationship. Similar conclusions can be drawn from the analysis of the ovendry weight of branches.

Let us analyze now the variation of the stratified ratio estimator of the bole ovendry weight (wood + bark). The variance ranges from 0.0007 to 1633.0545, corresponding to the smallest and largest sample trees, respectively (Table 4). As was the case with foliage and branch biomass, we have plotted the variance of the estimator of the bole biomass against tree dbh

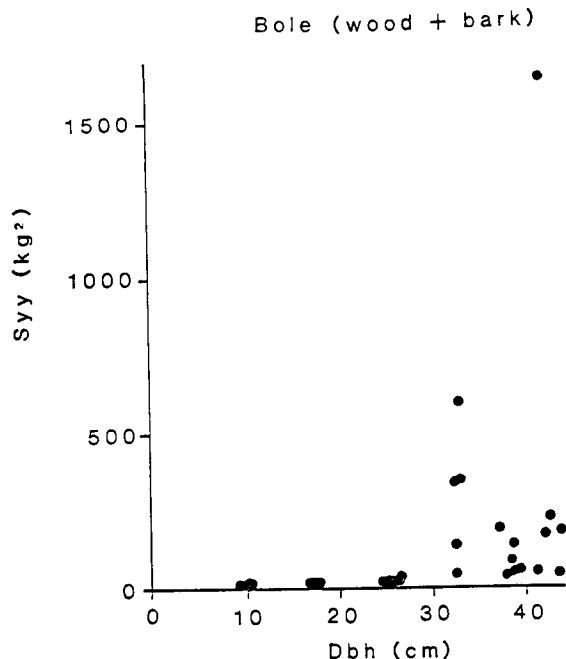


Figure 4. The variance of the stratified ratio estimator of the bole oven-dry weight (wood + bark) plotted against the tree diameter.

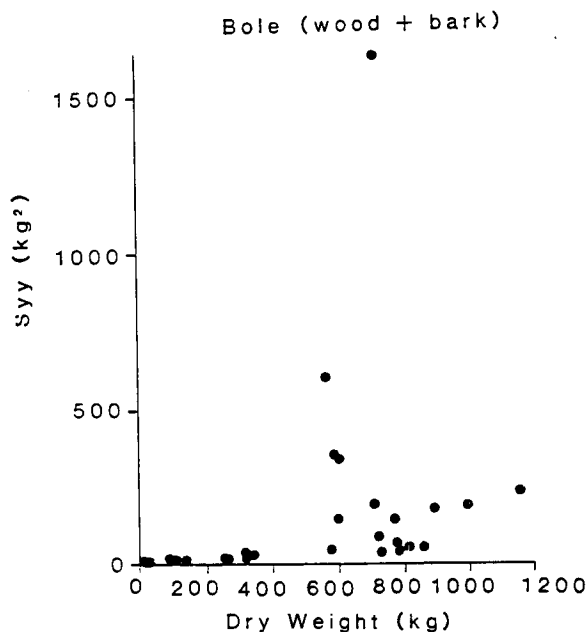


Figure 5. The variance of the stratified ratio estimator of the bole oven-dry weight (wood + bark) plotted against biomass.

(Figure 4) and against bole biomass itself (Figure 5). The graph of the standard error of the biomass estimator against the squared value of dbh (Figure 6) shows that the assumption that the variance estimator of bole biomass is approximately proportional to the fourth power of dbh is reasonable. Similar analyses for the bolewood and bole-bark components lead to the same conclusions.

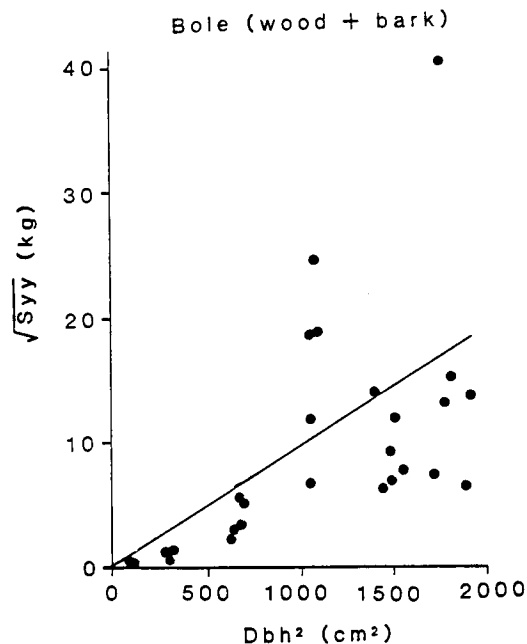


Figure 6. The standard error of the bole oven-dry weight (wood + bark) plotted against the squared value of tree diameter.

It is known that the ratio estimators are biased. As the estimator of the bole biomass is based on a sample size of nine disks, we would expect the bias to be small. Results from Table 4 verify this expectation; with very few exceptions the bias is much less than one percent of the biomass estimate. In terms of the standard error, however, the estimated bias is about seven percent for the first tree, about one percent for the fourteenth tree, and about 20 percent for the twenty-ninth tree. Whether a bias of this size can be tolerated is a question for management to answer. We feel that in terms of the resulting biomass table we construct, this bias is negligibly small. Similar conclusions can be reached from the corresponding analyses for the bolewood and bole-bark oven-dry weight data.

Concluding Remarks

It is generally expensive and time consuming to determine the biomass of sample trees by direct measurement. Most of the time, it is sufficiently precise to estimate the biomass by subsampling. This can be done in many ways. We have described the specific subsampling methods used in a study designed to evaluate the effects of some harvesting procedure on future site productivity.

The oven-dry weight of the tree foliage and branches was found by direct measurement. All the material was oven-dried and weighed. To facilitate the handling of this material, but much more importantly to obtain data that can be used later in the optimization of the subsampling method, the foliage and branches were divided into ten piles and the oven-drying, weighing and

recording were done separately by pile. By considering the estimation error when $m < 10$ piles only are sampled, one can make a decision as to how many sample piles would be sufficient to satisfy precision requirements.

The biomass of the main tree bole expressed as green weight can be determined easily by direct measurement; there is no need for subsampling. It requires the use of a field scale, and only when the tree is large requires sectioning, with each section being weighed separately. The measurement error is expected to be equal to zero if the weighing is carefully done. The biomass of the same bole, expressed as oven-dry weight, is however, much more time consuming and expensive to measure. For this reason it is necessary to determine oven-dry weight by subsampling.

The subsampling method used in this study was stratified random sampling and the estimator used was that of the stratified ratio type. The variance within strata was estimated for both green and oven-dry weight of the subsample units (disks) as well as their covariance and ratios. These statistics can be used in later study to decide the amount of subsampling that one may need to obtain estimates of the oven-dry weight of the main bole, should one decide to use stratified sampling and ratio estimators of the type used in this study.

It is in this sense that one must view the value of this paper. It provides information about the variation within sample trees. Quantitatively expressed, this information can be used in later studies to determine the amount of subsampling needed. Then, optimization models of the type suggested by Cunia (1986) can be used to find out at what point one should stop subsampling and add more sample trees so as to obtain maximum information for given costs of sampling or to minimize these costs and obtain estimates of desired precision.

Acknowledgements

This paper is based on research funded by the Research Foundation of the State of New York, the United States Department of Agriculture Forest Service, and the Department of Energy Grant No. 23-524 and Grant No. 23-708.

Literature Cited

- Briggs, R. D.; Cunia, T.; White, E. H.; and Yawney, H. W. Estimating sample tree biomass by subsampling. Syracuse, NY: SUNY College of Environmental Science and Forestry; Faculty of Forestry Misc. Publ. In Review. 1986.
- Briggs, R. D. Estimating biomass and nutrient removals from whole-tree and tree-length harvesting of a northern hardwood stand. Syracuse, NY: Ph.D. Thesis; 1985. 264 p.

Cunia, T. An optimization model for subsampling trees for biomass measurement. Paper presented at the Syracuse, NY, May 26-30, 1986 Workshop on Tree Biomass Regression Functions and Their Contribution to the Error of Forest Inventory Estimates. 1986.

UNBIASED ESTIMATION OF TOTAL TREE WEIGHT BY
THREE-STAGE SAMPLING WITH PROBABILITY
PROPORTIONAL TO SIZE

Harry T. Valentine, Timothy G. Gregoire, and
George M. Furnival

Research Forester, U.S. Forest Service, Hamden,
CT 06514, Assistant Professor of Forest
Biometrics, Virginia Polytechnic Institute &
State University, Blacksburg, VA 24061 and J.P.
Weyerhaeuser, Jr., Professor of Forest
Management, Yale University, New Haven, CT
06511.

A three-stage procedure for estimating the total tree weight in a plot employs discrete sampling with probability proportional to size, randomized branch sampling, and importance sampling--a continuous analogue of sampling with probability proportional to size. Published regression functions that predict tree weight from diameter breast height can be used to assign probabilities of selection to the trees for the first-stage sampling.

Introduction

In ecological or forestry research, a need often arises for an estimate of the total weight of the trees in a research or survey plot. The biomass literature is rife with regression functions that estimate tree weight from diameter. Consequently, an investigator might be tempted to measure the diameters of the trees in a plot, use published regression functions to estimate their individual weights, and then sum these weights to obtain the needed estimate of total tree weight. Unfortunately, the model-based estimate of total tree weight obtained by this simple procedure may be biased if the regression functions do not derive from the population to which they are applied. Bias also may arise if the population of current interest is a subpopulation of the parent population from which the regression functions derive. This bias may be large or small and of either sign, and the absolute magnitude of this bias does not decrease with increasing sample size.

In this paper, we describe a procedure that furnishes a sample-based unbiased estimate of the total weight of the trees in a plot. Initially, a sample of trees is selected at random with probability proportional to size (PPS). The weights of the trees in this sample are either measured or unbiasedly estimated in optional second and third stages of sampling. The second and third stages employ procedures for subsampling trees developed by Valentine et al. (1984).

Published regression functions may be used in connection with the first-stage PPS sampling. The probability of selection of a tree in the plot can be made proportional to weight as predicted by a regression function. Any regression function(s) could be used for this purpose. However, the precision of the PPS estimator depends on the degree of correlation between the predicted weights and the actual weights. Better correlation gives better precision.

First-Stage Sample

The use of PPS sampling for the estimation of the total tree weight in a plot has become feasible with the advent of programmable pocket calculators and field-worthy portable microcomputers. To proceed with PPS sampling, the N trees in a plot are numbered consecutively as their diameters are measured. Regression functions that predict tree weight from diameter are programmed into a calculator or computer and used to predict the weight of each tree. The predicted weight of the tree number i ($i = 1, 2, \dots, N$) is denoted by $\hat{w}_i$. The conventional model-based estimate of the total (fresh or dry) weight of the trees in the plot is obtained by summing the individual (fresh or dry) predicted weights. To obtain our sample-based estimate, we select a sample of n trees at random (with replacement) with probability proportional to predicted weight as follows:

The total predicted weight of the trees number 1 through i (C_i is accumulated for $i = 1, 2, \dots, N$):

$$C_i = \sum_{j=1}^i \hat{w}_j \quad i=1, 2, \dots, N \quad (1)$$

We also define $C_0 = 0$. To select the k th of n trees in our sample, a random number, u_k , is drawn from a uniform $[0, 1]$ distribution. The tree numbered i is selected as the k th tree in our sample if

$$C_{i-1} < u_k C_N \leq C_i \quad i=1, 2, \dots, N$$

We denote the tree (fresh and dry) weight of this tree numbered i by w_i . An unbiased estimate of the total tree weight in the plot, based on this k th tree in our sample, is

$$\tilde{w}_k = C_N w_i / \hat{w}_i \quad (2)$$

An unbiased estimate of the total tree weight based on a sample of n trees is

$$\tilde{W} = \sum_k^n \tilde{W}_k / n \quad (3)$$

and a sample-based estimator of the variance (var) of $\tilde{W}$ is

$$\text{var}(\tilde{W}) = \sum_k^n (\tilde{W}_k - \tilde{W})^2 / [n(n-1)] \quad n \geq 2 \quad (4)$$

Proofs of unbiasedness of PPS estimators can be found in most sampling texts (e.g., Cochran 1963, Sukhatme and Sukhatme 1970).

By performing PPS sampling as was described, we take advantage of the information extant in published regression functions, and we also can take advantage of a built-in safeguard against error. If there is a suspiciously large difference between $\tilde{W}$ and the conventional model-based estimate, C_N , then we should check for errors in measurement or entry of data into the calculator or computer.

There may be occasions when published regressions are not available for the species of interest, in which case other information might be used to advantage in a PPS sampling. For example, in 1927, C.D. Murray reported that the weight of a tree above a given cross section was proportional to the diameter (d) of that cross section raised to a power of 2.5. The report was based on 116 measurements of whole trees, branches, and leaves from nine tree species. In 1981, J. White cited 50 publications in which tree weight has been reported as proportional to diameter breast height (D) raised to a power (b). With just two exceptions, all values of b were between 2 and 3. White, like Murray 54 years before, concluded that 2.5 is a reasonable approximation of b for many tree species. White also noted that tree volume is correlated with $D^{2.5}$. Thus, in the absence of regression functions, we could sample with probability proportional to $D^{2.5}$ in pure stands, or with probability proportional to $gD^{2.5}$ in mixed stands, where g is an approximation of the (fresh or dry) weight per unit volume of the wood for a given species.

Second-Stage Sample

The PPS estimator furnishes an unbiased estimate of the total weight of the trees in the plot from measurements of the true weights of the trees in the PPS sample. If total fresh weight is the only parameter of interest, then the trees in the sample could be felled, weighed, and forgotten. Often, investigators are interested not only in total fresh weight but also in dry weight, mineral content, and volume. A subsampling procedure developed by Valentine et al. (1984) can be applied to the trees in the first-stage PPS sample to furnish

estimates of all of these parameters and their sampling errors. This procedure also eliminates the laborious task of weighing whole trees.

The subsampling procedure of Valentine et al. has two stages, the first stage employing randomized branch sampling (Jessen 1955), and the second stage employing importance sampling--a technique of Monte Carlo integration (e.g., Rubinstein 1981, Gregoire et al. 1985). The first and second stages of this subsampling procedure become, respectively, the second and third stages of the present procedure for estimating the total weight of the trees in a plot.

Randomized branch sampling is used to select a path that consists of connected branch segments (internodes), starting at the butt and ending at a terminal shoot of a tree. For operational purposes, a branch is defined as the entire stem system that develops from a single bud (lateral or terminal), whereas a segment is a part of branch between two consecutive nodes. We make no operational distinction between the main stem and side branches.

The first segment of any path extends from the butt (defined as the first node) to the second node. The second segment extends from the second node to a third node, and so on to the last (m th) segment of the path, a terminal shoot. The number of possible paths in a tree equals the number of terminal shoots.

Because the first segment of any path is fixed at the outset, its probability of selection (denoted by q_1) equals 1. To determine the remainder of our path we select one of the branches emanating from the second node at random with probability proportional to size, e.g., $d^{2.5}$. Selection of a branch fixes the second segment of the path. Its probability of selection is denoted by q_2 . The second segment is followed to a third node where a branch (and the third segment of the path) is selected with probability q_3 . This procedure is repeated until a terminal shoot is selected with probability q_m .

The probability q_j ($j=1,2,\dots,m$) is the conditional probability of selecting the j th segment of the path given that the path has reached the node to which the j th segment is attached. The unconditional probability of selection for the j th segment included in the path is:

$$Q_j = \prod_{r=1}^j q_r \quad (5)$$

Suppose that randomized branch sampling has been used to select a path on a tree numbered i in the first-stage sample. Denoting the weight of the j th segment in the path by b_{ij} , an unbiased estimate of the weight of this tree is (Valentine et al. 1984):

$$\bar{w}_i = \sum_j^m b_j Q_j^{-1} \quad (6)$$

This estimate could be substituted for w_i in (2) to furnish an unbiased estimate of the total weight of the trees in the plots.

Third-Stage Sample

The reciprocal of the unconditional probability of selection of the j th segment included in a path, Q_j^{-1} , is called an inflation factor. The product, $b_j Q_j^{-1}$, is called the inflated weight of the j th segment, and the sum of the inflated weights of the segments in a path--the unbiased estimator of tree weight--is called the inflated weight of the path. To measure the inflated weight of the path, we must separate and weigh the segments, some of which could be quite heavy. However, by employing the importance sampling procedure developed by Valentine et al. (1984), we can avoid weighing these segments. In fact, we can estimate the inflated weight of a path from the weight of a single disk. We also obtain a convenient specimen (the disk) for estimation of the inflated dry weight or mineral content of the path, which are, respectively, unbiased estimators of the dry weight or mineral content of the tree.

Importance sampling is a continuous analogue of PPS sampling and is used for estimating integrals. The inflated path weight, w_i , can be defined as the integral:

$$\bar{w}_i = \int_0^\lambda Y(L) dL \quad 0 \leq L \leq \lambda \quad (7)$$

where $Y(L)$ is the inflated weight per unit length at point L along the path, and λ is the length of the path. (Any quantity measured at a point on the j th segment of the path is inflated by dividing the quantity by Q_j .)

By analogy to discrete PPS sampling, our task is to select a sample of points along the path independently and at random with probability as nearly proportional to inflated weight per unit length as possible. As was noted, tree weight is highly correlated with volume. Therefore, inflated weight per unit length, $Y(L)$, is highly correlated with inflated cross-sectional area $[A(L)]$, because cross-sectional area is volume per unit length. Although $A(L)$ is unknown, it is easy to build an *ad hoc* proxy function that accurately approximates $kA(L)$ (where k is an arbitrary constant).

For example, Valentine et al. (1984) described how to build a segmented-linear function, $S(L)$, that interpolates inflated d^2 measured at $L=0$, $L=\lambda$, and at arbitrary points in between. The resultant function,

$S(L)$, approximates $1.27A(L)$, is positive in the range $0 \leq L \leq \lambda$, and integrates to give

$$V(\lambda) = \int_0^\lambda S(L) dL \quad (8)$$

Selection of the z th of M points along the path at random with probability proportional to $S(\theta_z)$ involves finding the root, θ_z , of

$$V(\theta_z) - u_z V(\lambda) = 0 \\ 0 \leq \theta_z \leq \lambda; z=1, 2, \dots, M \quad (9)$$

where u_z is a uniform $[0,1]$ random number. An unbiased estimate of w_i based on a sample of M points is

$$\bar{w}_i = [V(\lambda)/M] \sum_{z=1}^M Y(\theta_z)/S(\theta_z) \quad (10)$$

Substitution of $\bar{w}_i$ for w_i in (2) gives an unbiased estimate of total weight of trees in the plot.

In application we find that inflated weight per unit length can not be measured at a point, θ_z . Instead, we must measure weight per unit length of a disk centered at θ_z , thereby creating a trivial technical bias. However, it is possible to measure inflated cross-sectional area at θ_z , and substitution of this quantity for $Y(\theta_z)$ in (10) gives an unbiased estimate of the volume of the tree that can be substituted for w_i in (2) to give an unbiased estimate of the total aboveground woody volume of the trees in the plot.

The subsampling procedure involving randomized branch sampling and importance sampling is described in detail by Valentine et al. (1984). Results of a field test are reported by Valentine et al. (1984) and in somewhat greater detail by Tritton et al. (1983).

Discussion

Our procedure for estimation of total tree weight on a plot requires some on-site calculations, but very little office work. We have attempted to design a procedure whereby an investigator can enter a plot with a programmable calculator or portable computer and leave with an essentially unbiased estimate of fresh weight (or volume) and an estimate of sampling error, if desired. If the parameter of interest is dry weight or chemical constituency, then an investigator can leave the plot with specimens for drying or analysis and their respective inflation factors.

Invariably, complications arise in any forest sampling. In a randomized branch sampling of a broad-leaved tree, we may have to

deal with epicormic branches. If there are not too many, they can be removed and weighed separately. If there are a lot, we can ignore them while selecting the path by randomized branch sampling. An estimate of the weight of the epicormic branches on the tree can be calculated as the sum of the inflated weights of the epicormic branches on the segments of the path. Dead branches can be treated the same way.

In a randomized branch sampling of a coniferous tree, we may find a whorl containing a large number of small side branches at the j th node on the main stem. Consequently, assignment of selection probabilities based on diameter measurements becomes tedious. We can attempt to avoid measuring these small branches by performing the selection of the path segment in two stages. Let d_a and d_b denote, respectively, the main-stem diameter above and below the j th node. Based on Murray's (1927) observations, the sum of $d^{2.5}$ for the branches in the whorl plus $d^{2.5}$ should be approximately equal to $d_b^{2.5}$. If $d_a < d_b$ as expected, we can randomly select the main stem with probability $q_j = (d_a/d_b)^{2.5}$ or the side branches collectively with probability $1 - (d_a/d_b)^{2.5}$. The side branches need not be measured if the main stem is selected. If the collection of side branches is selected, we select one of the side branches at random with probability q'_j whence $q_j = q'_j[1 - (d_a/d_b)^{2.5}]$.

The techniques of importance sampling should be applicable for estimation of many quantities of interest to foresters and ecologists. Since its seminal use in a forestry problem by Valentine et al. (1984), importance sampling has been employed in procedures for estimation of bole volume (Gregoire et al. 1985, 1986), bole-volume increment (Valentine et al. 1986), and log volume (Furnival et al. 1986). Each of these procedures eliminates the bias that arises when a volume formula is used for estimation in a population other than that from which the formula derives.

Literature Cited

- Cochran, W. G. Sampling techniques. New York: John Wiley & Sons; 1963. 413 p.
- Furnival, T. G.; Valentine, H. T.; Gregoire, T. G. 1986. Estimation of log volume by importance sampling. For. Sci. 32(4): 1073-1078.
- Gregoire, T. G.; Valentine, H. T.; Furnival, G. M. Estimation of volume with a taper function: A sampling approach. In: Use of auxiliary information in natural resource inventories; 1985 October 1-2; Blacksburg, VA. Soc. Am. For. Publ. No. SAF86-01. 1985: 164-170.
- Gregoire, T. G.; Valentine, H. T.; Furnival, G. M. 1986. Estimation of bole volume by importance sampling. Can. J. For. Res. 16: 554-557.
- Jessen, R. J. Determining the fruit count on a tree by randomized branch sampling. Biometrics II: 99-109; 1955.
- Murray, C. D. A relationship between circumference and weight in trees and its bearing on branching angles. J. Gen. Physiol. 10: 725-729; 1927.
- Rubinstein, R. Y. Simulation and the Monte Carlo method. New York: John Wiley & Sons; 1981. 278 p.
- Sukhatme, P. V.; Sukhatme, B. V. Sampling theory surveys with applications. 2nd rev. ed. Ames, IA: Iowa State University Press; 1970. 452 p.
- Tritton, L. M.; Valentine, H. T.; Furnival, G. M. A new procedure for estimation of tree biomass and nutrient content. In: Mesures des biomasses et des accroissements forestiers; 1983 October 3-7; Orleans, France. Les Colloques de l'INRA, No. 19; 1983: 335-341.
- Valentine, H. T.; Tritton, L. M.; Furnival, G. M. Subsampling trees for biomass, volume, or mineral content. For. Sci. 30: 673-681; 1984.
- Valentine, H. T.; Gregoire, T. G.; Furnival, G. M.; Solomon, D. S. Unbiased estimation of bole increment. In: Environmental influences on tree and stand increment. IUFRO (S4.01-06); 1985 September 23-27; Durham, NH. [in press].
- White, J. The allometric interpretation of the self-thinning rule. J. Theor. Biol. 89: 475-500; 1981.

MEASUREMENT ERRORS IN FOREST

BIOMASS ESTIMATION

Daniel AUCLAIR

INRA, Centre de Recherches d'Orléans,
Station de Sylviculture
ARDON - 45160 OLIVET (France)

When sampling for biomass regressions, one source of error can come from the error made on the independent variables. The order of magnitude of this error has been estimated on several variables : girth at breast height, tree height, relative water content of samples, volume of samples, ... Their contribution to the error on the estimation is discussed.

Introduction

For estimating biomass of sample trees, to be subsequently applied to regressions in order to determine prediction equations, many procedures have been developed by various forest scientists. These have been well described by YOUNG (1976), PARDE (1980), SATOO and MADGWICK (1982), and one of the annual meetings of IUFRO's S4.01-00 subject group on "Mensuration, Growth and Yield" was devoted to "Problems in forest biomass mensuration and growth and yield studies" (AUCLAIR, 1983).

Most methods use diameter (or girth) at breast height, or tree height, or a combination of both, as independent variables for biomass regressions. The weight of sample trees can be estimated by several methods, including either a conversion factor from fresh weight to dry weight or a conversion factor from volume to dry weight. This estimation can be done on the whole tree, or if it is too big on subsamples (branches) which may be used once more for establishing regressions for individual tree biomass estimation.

At each of the steps in the process of determining biomass regression equations, some measurement errors occur. These may be random errors, often assumed to be normal, or systematic errors.

In order to obtain the best possible estimations, the present study was undertaken to determine the various contributions of measurement errors in the procedures used in our laboratory for estimating biomass. The practical aim was to minimize the most important errors, at a "reasonable" cost.

Most of our studies concern trees treated as coppice : traditional coppice with 20- to 30-year rotations, or short-rotation coppice going down to annual rotations. These are therefore small trees, whose biomass can be determined

by estimating the fresh weight of the whole tree, most often separated in several components, and collecting samples for water content.

However in some cases we have also estimated biomass of larger trees, from traditional high forest or coppice with standards, where the stem volume was estimated geometrically, and samples were collected to convert volume into dry weight.

This paper concerns estimations of measurement errors on :

- girth at breast height
- total tree height
- water content of samples
- weight of samples
- volume of samples
- the use of basic density for biomass estimation.

The data comes either from original mensurations collected especially for this purpose, or from other studies by AUCLAIR and METAYER (1980) or by BISCH (1985).

1 - Girth and Tree Height

a - Method : On twenty Chestnut trees (*Castanea sativa* Mill.) situated in a traditional coppice stand of the INRA estate, aged from four to thirty five years, the girth at breast height was measured, using a metal tape graduated in millimetres. Fifteen different persons, forest research professionals, technical staff, and also some non-professionals, measured the same trees, giving the girth with an accuracy of 1 cm.

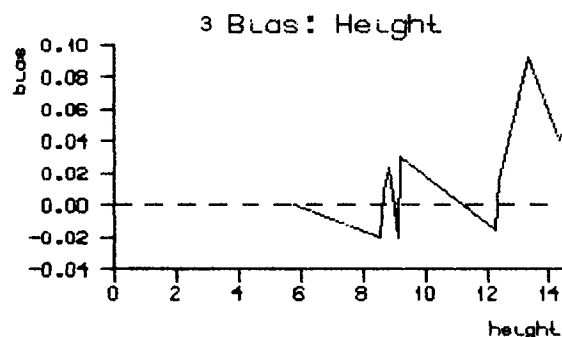
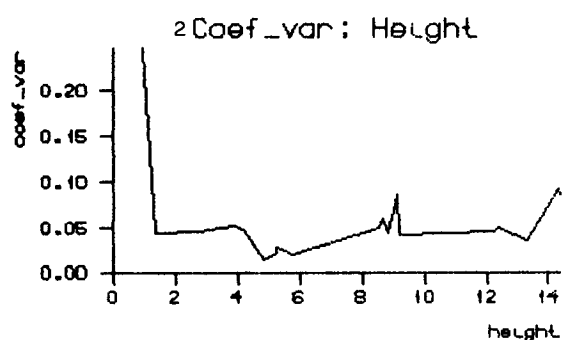
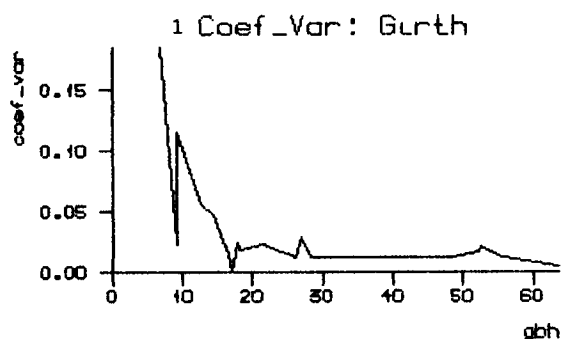
On the same trees the total tree height was estimated :

- 1) with a graduated rod on trees smaller than 8 m., with an accuracy of 1 cm.
- 2) with a Blume-Leiss hypsometer for trees 8 to 15 m high, with a 10 cm accuracy. On these trees a possible bias was estimated by one single measurement with the graduated rod.

To verify whether or not there was any large difference between operators, we used the non-parametric test of Kendall's coefficient of concordance. In all cases it showed no difference, at a very high degree of significance, both for girth and height measurements.

b - Results : The coefficients of variation and the bias for height measurements are shown in Figures 1 to 3, plotted against the mean of either girth-at-breast-height, or total height.

The two high values of the coefficient of variation for smaller girths (Fig. 1) can be explained by the data collected by one operator who overestimated the girth by 2 cm compared to the mean. These figures however still remain around 0.10 if this operator is excluded, but this corresponds to absolute differences of ± 1 cm.



Figures 1 to 3 : coefficient of variation of girth or height plotted against girth (cm) or height (m), and percentage of bias for tree height.

The very high value for the smallest tree in Figure 2 comes from an error in transcription of the data where a tree 1.00 m high was noted 0.30 m. If this data is excluded the coefficient of variation becomes 0.09, which corresponds to an absolute difference of ± 15 cm.

In Figure 1, after 20 cm the coefficient of variation for girth smooths out at a value under 0.02, whereas that for height (Fig. 2) seems to increase slightly with increasing height, but remains around 0.05.

The bias estimated on height measurements (Fig. 3) seems to increase with height, but it is much more correlated with the fact that the crown of a tree is more or less visible : a "visibility index" was applied to each of the measured trees, and the coefficient of variation and bias for height were compared to this index :

Visibility	Coef-var (%)	bias (%)
good	5.7	- 0.3
medium	3.6	+ 0.7
poor	7.2	+ 5.4

The height of poorly visible trees seems to be overestimated, but no conclusion can be reached about the coefficient of variation.

c - Conclusions : The variation in girth and height measurements is quite high for very small trees, however their contribution to the total biomass of a stand is usually rather small. The variation for girth levels out quickly for larger trees, but variations in height measurement remain at quite a high value, and may increase for the bigger trees.

Height measurements are very time-consuming, and it would be rather costly to try to increase precision by improving the accuracy of these measurements for all the trees. We therefore suggest to focus on accurate "top height" measurements and use the "top height" of a stand as an independent variable for biomass regressions. This improves the precision compared to regressions with girth or diameter as the only independent variable (PAGES, 1985).

The bias which was estimated here may seem unimportant, as the same bias should be obtained when establishing the regression and when using it for prediction. However the sampling should take into account the fact that the tree crown is more or less visible, as a "poor" visibility seems to provoke an overestimation.

2 - Water content of samples

a - Variation within a tree : In a previous report, AUCLAIR and METAYER (1980) studied the water content of three series of samples on trees aged 25 to 40 (Birch, Oak, and Hornbeam) divided in five diameter classes : 0-2.5 cm, 2.5-4 cm, 4-7 cm, 7-15 cm, + 15 cm. The results (Fig. 4) show that the water content increases in the smaller components.

Several authors have tried to use only one single mean sample for estimating total dry biomass, but as the proportion of different components varies between species, or between trees (Fig. 5), it seems most important to apply the appropriate conversion factor to each of the components.

Water content

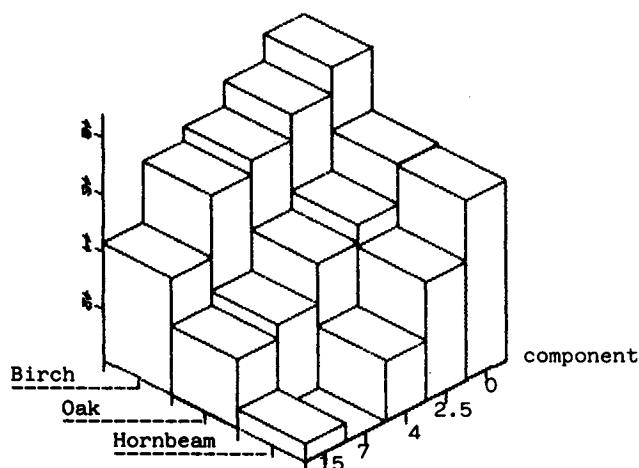


Figure 4 : Water content (percent fresh weight) for four components in Birch, Oak, and Hornbeam.

Biomass distribution : Birch

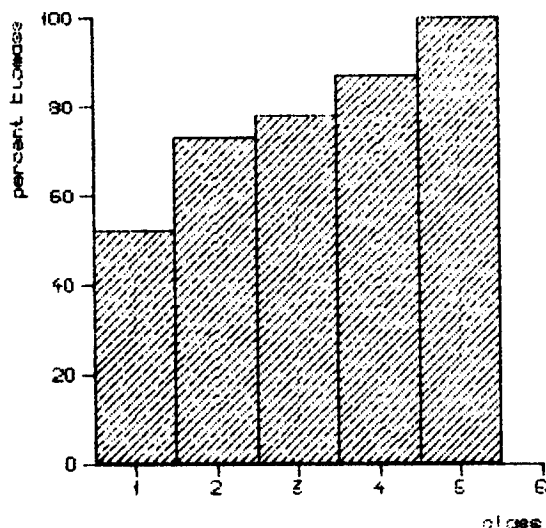


Figure 5 : Percentage of biomass in each class for 25-year Birch coppice :

- 1 : > 7 cm
- 2 : "stacked wood"
- 3 : > 4 cm
- 4 : > 2.5 cm
- 5 : total

b - Drying temperature : For drying samples to estimate water content, the usual procedure is to leave them in an oven until their apparent weight becomes constant. This usually lasts two to five days.

However, some authors suggest using an oven temperature of 60 or 65 degrees C., others 85 de-

grees, and others 105 degrees. The lowest temperature is used mostly for drying leaves, or when nutrients are analyzed ; at the highest temperature all the water evaporates but some volatile compounds may also disappear.

On a series of samples, divided in four diameter classes, we tested the relative difference in apparent dry weight after drying at 65, 85, or 105 degrees :

Diameter	$\frac{W_{65}-W_{85}}{W_{65}} \%$	$\frac{W_{85}-W_{105}}{W_{85}} \%$
0-2.5 cm	3.9	1.3
2.5-4 cm	4.5	2.4
4-7 cm	4.7	6.4
+ 7 cm	3	8

In all cases the samples still loose some weight between 85° and 105°. The smaller diameter classes (< 4 cm) loose less than 2.5 percent, but the larger ones loose between 5 and 10 percent.

It therefore seems quite difficult to give a single conversion factor from the dry weight at one oven temperature to the dry weight at another, for the same reason as above, as the distribution of diameter classes inside each tree is most variable.

c - Water loss of samples : In the process of biomass estimation, it is often impossible to estimate the fresh weight of samples immediately at the same time as the whole tree or compartment biomass estimation. The type of balance used is often not sufficiently precise for weighing small samples. If the fresh weight is to be estimated at the laboratory, we often keep the samples in hermetically closed plastic bags.

To estimate a possible error due to this practice, we measured the fresh weight of samples in the field, and after keeping them in plastic bags for 1, 2, 3, or 10 days. Figure 6 shows the distribution of relative differences at the four sampling dates.

After one day, most samples lose 1 or 2 percent of their fresh weight, but after two or three days the weight loss can reach 6 percent. After ten days the distribution of error smooths out, the weight loss varies from 1 to 8 percent, and in one case 10 percent.

As in the study concerning different drying temperatures, it is not possible here to give a single correction factor due to the delay in weighing samples. However, if the samples are weighed inside the plastic bag still hermetically closed, their weight loss is compensated by water kept inside the bag, and the error is negligible.

number of
samples

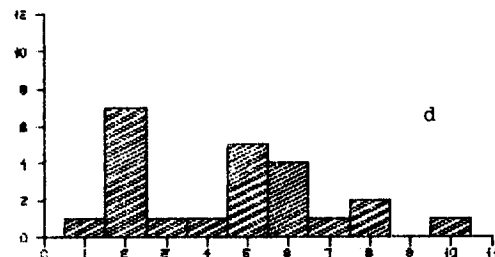
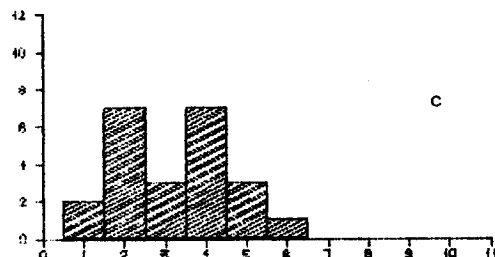
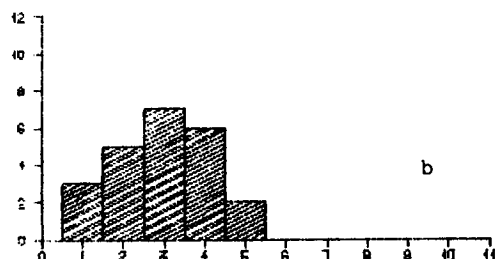
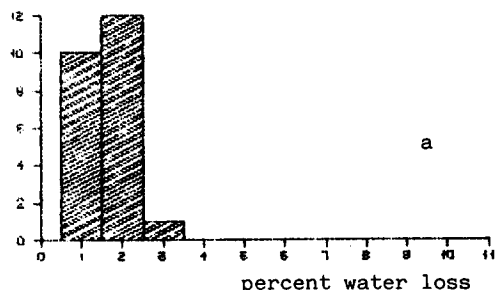


Figure 6 : Histogram of water loss of samples left in plastic bag for :

- a : 1 day
- b : 2 days
- c : 3 days
- d : 10 days

3 - Weight of samples

For biomass estimation in the field, we usually weigh tree compartments on a very precise balance, whose scale is graduated every 20 grams, up to 60 kg.

Although this balance is sent every year to have a complete check-up, we have verified in various occasions the values indicated. One exam-

ple is given in figure 7. It shows the absolute value of relative error, plotted against total weight. All through the scale there is a slight underestimation. For lower values the underestimation is 40 g, and for higher values it can reach 80 g. This leads to quite a small relative error, less than one percent, but it is however possible to give a correction factor for this systematic bias if such curves as that in figure 7 are regularly verified.

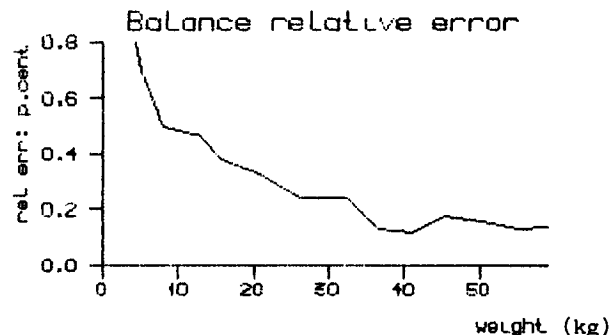


Figure 7 : Relative error (x 100) of balance used for biomass estimation.

4 - Volume of samples

For estimating biomass of very large trees, a method often applied is to estimate the bole volume, and use a conversion factor estimated from a sample, or several samples taken from various levels on the bole.

Such a study was undertaken by BISCH (1986), who showed that there is often a bias in the estimation of the volume of disks, due to the fact that the Archimedean principle is often incorrectly used. The fact that the quantity of water absorbed by a sample during its immersion is not taken into account, produces a systematic underestimation of the volume. On oak disks he showed that the mean bias was above 1 percent, and can reach 2 percent. This seems highly correlated with the bark volume of the samples.

However, as the bole volume of a tree is estimated geometrically, it seems more important to also estimate geometrically the volume of samples. This was done by measuring the disk girth and the disk width as a mean of four different measurements, with a metal tape.

The geometric volume gives an overestimation compared to "real volume" estimated by weight of water displacement. The relative bias is 5 percent $\pm$ 0.4 percent, estimated on 74 disks from two different oak species (*Q. robur* and *Q. petraea*). The author noticed no difference between species, nor between silvicultural treatments.

5 - The use of basic density

In his study of biomass distribution, BISCH (1985) estimated bole volume with two different methods, and estimated wood basic density on four

disks from the base of the tree, 4 m, 8 m, and 12 m upwards.

Volume 1 is estimated with the traditional method of the french national forest survey, dividing the bole in two parts : under, and above 2.6 m, and measuring the girth at the middle of each of these sections.

Volume 2 is estimated by measuring diameters every meter along the bole.

Four different total stem biomass estimations were made :

B1 : volume 1 and a mean of the basic densities of the four samples ;

B2 : volume 2 and a mean basic density ;

B3 : volume 1, and basic density from the base of the tree being applied to the first section, a mean of the others to the above section ;

B4 : volume 2, divided in four compartments, 0-1 m, 1-6 m, 6-10 m, > 10 m.

The mean differences are the following :

$$\frac{B1 - B4}{B4} \times 100 = - 2.13$$

$$\frac{B2 - B4}{B4} \times 100 = + 0.59$$

$$\frac{B3 - B4}{B4} \times 100 = - 2.53$$

The use of the "traditional" forest survey volume underestimates the total stem biomass. The basic density varies up the stem (Fig. 8), and using the mean of four disks seems to give a reasonably good estimation. The use of one single sample for basic density estimation seems insufficient.

height (m)

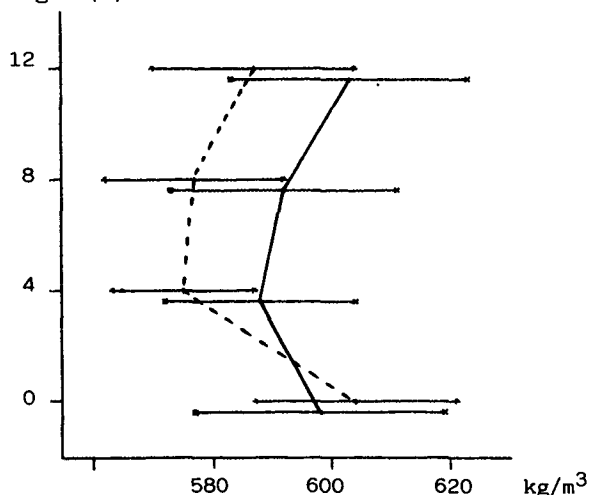


Figure 8 : Mean basic density of samples of Oak at different levels in the stem. Horizontal bars indicate confidence intervals at 5 % :

— standards from coppice with standards
 ---- trees from high forest

Conclusions

In our studies on biomass estimation, we have attempted to minimize measurement errors, and especially bias. Several phases in the process of biomass estimation have been studied here, and suggestions can be given to improve the accuracy of measurements. However, the quality/cost ratio may in some cases become very low, and some errors are difficult to minimize.

The same type of study has already been undertaken for volume estimation (SCHMID-HAAS et al., 1980 ; WINZELER, 1986) suggesting the best type of material to use or measurements to be done. The present study brings a contribution to measurement errors on biomass estimation.

The effect of various types of error on the results of regression utilization has been studied by GERTNER and DZIALOWY (1984) and GERTNER (1985), who showed that for example a 10 % bias in DBH may cause a 25 % error in predicted basal area.

This type of bias is quite in the same order of magnitude as the data presented here. It must however be outlined that the present study concerns mostly traditional coppice, with very bad stem form, most often on very old root systems. In other cases, for example conifers who have much straighter stems, or with more homogeneous plant material, as genetically controlled plantations, the errors may be reduced.

Literature cited

- AUCLAIR, Daniel (Ed.). Problems in forest biomass mensuration and growth and yield studies. Proceedings, IUFRO S4.01-00 ; 1983 October 3-7 ; Orléans (FRA). Les colloques de l'INRA, 19 ; I.N.R.A., Paris ; 1983. 356 p.
- AUCLAIR, Daniel ; METAYER, Sylvie. Méthodologie de l'évaluation de la biomasse aérienne sur pied et de la production en biomasse des taillis. Acta oecol., Oecol. appl. 1 (4) : 357-377 ; 1980.
- BISCH, Jean-Luc. Influence du traitement sylvicole (futaie et taillis sous futaie) sur la répartition de la biomasse dans le Chêne en région Centre. INRA, Station de Sylviculture, Orléans (FRA) ; 85/42 ; 1985. 26 p.
- BISCH, Jean-Luc. Le taux de conversion de volumes de bois frais en biomasse : Amélioration de ses méthodes d'estimation. Ann. Sci. for. 43 (3) : in press ; 1986.
- GERTNER, George Z. Regressor variable errors and the estimation and prediction with linear and nonlinear models. October 1985. 15 p.
- GERTNER, George Z. ; DZIALOWY, Paul J. Effects of measurement errors on an individual tree-based growth projection system. Can J. For. Res. 14 : 311-316 ; 1984.

- PAGES, Loïc. Le taillis de Robinier du Val de Loire - croissance, biomasse, régénération. Université de Paris-Sud (Orsay) ; thèse (Dr. es Sci.) ; 1985. 74 p.
- PARDE, Jean. Forest biomass. For. Abstr. 41 (8) : 343-362 ; 1980.
- SATOO, T. ; MADGWICK, H. A. I. Forest biomass. The Hague : Martinus Nijhoff/Dr. W. Junk ; 1982. 152 p.
- SCHMID-HAAS, P. ; WERNER, J. ; BAUMANN, E. Fehler bei der Rundholzmessung. Schweiz. Z. Forstwes. 131 (9) : 801-820 ; 1980.
- WINZELER, Klemens. Volumenbestimmung durch Messungen am stehenden Baum. Eidg. Anst. forstl. Versuchswes., Mitt. 62 (1) ; 1986. 99 p.
- YOUNG, Harold E. (Ed.). Oslo biomass studies. Proceedings, IUFRO S4.01-00 ; 1976 June 22 ; Oslo. College of life sciences and agriculture, University of Maine at Orono, ME ; 1976. 302 p.

RESEARCH PAPERS

Biomass of Forest Understory Vegetation

Moderator: Delbert R. Mead

BIOMASS-DIMENSION RELATIONSHIPS OF
UNDERSTORY VEGETATION IN RELATION TO SITE
AND STAND AGE.

Paul B. Alaback

Research Ecologist, USDA Forest Service,
Pacific Northwest Research Station,
Forestry Sciences Laboratory, P.O. Box 909,
Juneau, AK 99802.

Three woody shrubs, two ferns, and three herb species in a wide range of stand types in southeast Alaska and the Pacific Northwest were examined. Biomass regression equations were developed for each of the species in recent clearcuts, mature second growth forests, old growth, and, for one species, in three thinning treatments. All but one of the fern equations were significantly different between age classes ($P < .05$). Herb biomass per unit cover was greatest in clearcuts and least in second growth and old growth. Woody biomass per stem size-class of shrubs was greatest in clearcuts and least in old growth. Leaf and twig production had the opposite relation. Differences between shrub biomass equations were generally related to shrub density and openness of the overstory canopy; changes in herb biomass were mostly related to openness of overstory canopy. Thinning experiments confirmed these basic patterns of change in biomass-dimension relations with overstory density. Use of equations from other regions gave significant errors in estimates of biomass compared with estimates made from equations developed in southeast Alaska. Equations developed for the same species in geographic regions with similar overstory or climate gave no more accurate estimates of biomass for southeast Alaska than equations developed in distant regions or from regions with contrasting climates. Developing equations for local populations or for specific stand structures is recommended for forest understory species.

Introduction

Understory biomass has been used to contrast plant community habitat types, to distinguish subtleties in forest microclimate, and to indicate quality of wildlife habitat. For more than two decades, regression equations have been used to predict understory biomass, principally for production, nutrient cycling, and studies of wildlife habitat (Alaback 1982, Ohmann and Grigal 1985, Tappeiner and John 1973, Whittaker 1961, Yarie 1980). Almost all of the equations developed were based on sampling from one study area or from several study plots

within a general study area. Most authors stressed the requirement that their equations be used on similar sites, but equations are often used in sites significantly different from the original sampling areas because of poor documentation of the original site conditions and locations.

Improper use of biomass equations developed from dimension analysis has been furthered by compilations of libraries of regression equations, usually with little information on how they were developed, what their limitations are, how they might be best applied, and what magnitude of error or bias is likely if they are applied to different sites (Gholz et al. 1979, Smith and Brand 1983, Stanek and State 1978). Some users of these libraries suggest an average of coefficients be used where more than one equation is available for a species, which may compound problems of potential error and bias (Ohmann and Grigal 1985, Pastor et al. 1984).

In this study, equations are presented for understory species growing in different sites in southeast Alaska and in comparable forest types in British Columbia, Montana, Minnesota, and Oregon. Objectives were: (1) to estimate the magnitude of bias in biomass estimates arising from using equations developed from different stand age-classes and overstory structures in southeast Alaska (2) to document the general pattern of variation in biomass-dimension relations for a given species across a gradient of overstory density and understory productivity (3) to examine the bias in biomass estimates made from equations developed in other geographic regions for a given species.

Methods

Sites with relatively homogeneous overstory structure, gently sloping terrain, and well-drained to relatively well-drained soils were selected in southeast Alaska (Alaback 1982, 1986). In each site, samples were randomly selected by plant size-class so that samples were more or less equally distributed across the full range of plant-size classes. Sampling was originally done during the summers of 1977, 1978 and 1979 coincident with a large-scale study of postlogging succession (Alaback 1982). Sampling was done between July 1 and September 1 of each year when all species had attained peak biomass (Alaback 1986). In the summers of 1982 to 1984 additional sampling was done to develop equations for new species, and to provide for sufficient samples in each of three stand-age classes to allow for direct comparison of equations. In addition, one site that had been precommercially thinned 4 years before biomass sampling was

studied. Biomass samples were taken from three of the thinning treatments (control, 8-foot spacing, and 16-foot spacing (Alaback 1984)).

Non-destructive measurements were taken on individual plants before destructive sampling. Samples were separated into components in the field, then oven-dried at 70 °C for 24-48 hours. Samples were weighed to the nearest 0.001 gram. Independent variables that were easy to measure in the field and were thought to bear a precise relation to biomass were selected: for shrubs, basal diameter; for herbs, percent ground cover on 0.1 m² plots; for ferns, frond length was used. Both ln-ln transformed and untransformed linear models were applied to the data for each species and biomass component. Models with the highest correlation coefficient and the least mean square error were selected as the best fit. Correction for bias was used for the ln-ln transformed data (Baskerville 1972).

Equations were compared by use of successive analyses of covariance with dummy variables (Cunia 1973). For each species, a regression model that included slopes and intercepts for all sites and treatments was used. All possible equations for a species were tested for differences in slopes and intercepts. If differences were significant then all possible pairs of equations were compared to determine which were significantly different. Pairs of equations with calculated F statistics greater than F at P = 0.05 were judged to be significantly different. Equations between regions for which original data were not always available were compared by calculating biomass estimates at an arbitrary value of the independent variable, usually at the upper limit of sampling where the greatest differences occurred.

Nomenclature for trees, shrubs, and herbs follows Hulten (1968).

Results and discussion

Herbs

Equations for specific age-classes or stand types tended to have coefficients of determination and mean squared error within the range of or better than regional biomass equations (Alaback 1986, Table 1). Herb cover equations were the most precise; shrub foliage and twig production were generally least precise. The independent variable for fern equations was frond length rather than percent ground cover because of the large variation in frond orientation with frond size or plant age. For both *Athyrium* and *Dryopteris*,

regression coefficients for open clearcuts and old growth were similar and are probably consistent over a wide geographic range. The primitive growth form and the lack of variation in branching or canopy structure probably contribute to the relative site independence of these equations.

For each of the three herb species, equations based on percent cover gave similar magnitudes and direction of bias for stand age-class (Figures 1-3). In recent clearcuts, herbs are generally taller and have denser foliage. In closed forests where light and temperature may be limiting factors, herb biomass is as little as 20-25 percent of biomass per unit cover in clearcuts at maximum cover values. Plants in closed forests are generally shorter with sparser canopies. The biomass equations also predict that plants growing in second-growth forests have slightly less biomass per unit cover than plants in old growth (P < 0.05, Table 2), consistent with the hypothesis of light limitation.

Shrubs

All three shrub species showed a similar pattern of change in biomass-dimension relations across the three age-classes (Table 1). In clearcuts, woody biomass was much greater per unit stem diameter than in the closed forest types (Figure 4). Leaf and twig production per unit stem diameter was much less than in the old-growth sites (Figure 5). Although the largest differences were between equations for recent clearcuts and closed forest types, significant differences were detected between biomass equations for old- and second-growth sites (Table 2).

In general, wood production per unit stem diameter was proportional to solar radiation, plant vigor, and shrub density, but leaf and twig production per unit stem diameter responded inversely to shrub density and incident radiation. Shrubs growing in open areas with plenty of light thus put more energy into height growth and stem development than shrubs growing in closed forests, which presumably had less competition with other understory plants and more need for expanded leaf area and canopy development to capture limited incident solar radiation (Gholz 1978). Equations from second-growth stands were probably influenced by the low vigor of residual plants, slowly being shaded out by overstory species (Alaback 1982). Many of these plants were tall with sparse canopies and were restricted to growing in small gaps in the overstory canopy. The dramatic change in absolute biomass of each component and proportional allocation between components argues strongly for reporting shrub biomass by component rather than simply as total aboveground biomass, which confounds these differences.

Table 1. Biomass equations for species in southeast Alaska growing in different stand types or age classes. All equations are significantly different at $P < .05$ using the techniques of Cunia (1973).

Species	a	b	r^2	mse	n	model**	range
HERBS TOTAL BIOMASS							
<u>Athyrium felix-femina:</u> *							
Clearcut	-3.226	0.047	.967	0.061	18	LL	18-105
Second growth	-2.568	0.034	.902	0.200	15	LL	32-147
<u>Cornus canadensis:</u>							
Clearcut	-2.665	1.278	.951	0.188	8	LC	2-100
Second growth (1)	-3.694	1.183	.834	0.205	17	C	3-60
Second growth (2)	0.020	0.050	.820	0.145	17	C	3-60
<u>Coptis aspleniifolia:</u>							
Clearcut	-2.567	0.293	.731	31.202	11	C	2-95
Second growth	-0.060	0.049	.955	0.009	8	C	2-25
Old growth	0.054	0.032	.972	0.006	7	C	5-40
<u>Dryopteris austriaca:</u>							
Clearcut	-9.036	2.286	.766	0.545	25	LL	9-95
Second growth	-8.841	2.178	.942	0.182	109	LL	4-95
<u>Rubus pedatus:</u>							
Clearcut	-0.578	0.100	.963	0.182	11	C	5-68
Second growth	0.103	0.028	.939	0.038	13	C	1-90
Old growth	0.244	0.011	.674	0.033	10	C	1-60
SHRUBS TWIG AND LEAF BIOMASS							
<u>Oplopanax horridum:</u> ***							
Clearcut	-6.314	3.684	.969	0.121	5	LD	4-27
Second growth	-3.205	2.128	.639	0.639	11	LD	5-30
Old growth	-5.488	3.579	.957	0.116	10	LD	5-20
<u>Rubus spectabilis:</u>							
Control	-3.754	2.842	.791	0.466	20	LD	2-15
Light thinning	-2.141	2.122	.690	0.322	11	LD	4-15
Heavy thinning	-1.114	1.923	.450	0.602	19	LD	6-25
<u>Vaccinium alaskaense:</u>							
Clearcut	-2.684	1.820	.660	0.816	82	LD	1-26
Second growth	-2.961	1.841	.810	0.433	75	LD	1-22
Old growth	-4.301	2.803	.730	1.262	7	LD	4-23
SHRUBS WOODY BIOMASS							
<u>R. spectabilis:</u>							
Control	-6.710	4.523	.914	0.295	10	LD	2-15
Light thinning	-2.680	2.636	.919	0.102	10	LD	4-15
Heavy thinning	-3.632	2.930	.733	0.402	20	LD	6-25
<u>V. alaskaense:</u>							
Clearcut	-2.247	2.553	.519	2.132	52	LD	1-26
Second growth	-1.612	1.701	.561	1.194	40	LD	1-22
Old growth	-3.231	3.210	.793	1.164	7	LD	4-23

* clearcut = 0-25 years after logging, second-growth = 26-200 years, old growth > 200 years.

** LL: $Y = \exp(a+b \cdot \ln(\text{length, cm}))$
 C: $Y \text{ (g/1000 cm}^2\text{)} = a+b \cdot (\text{percent cover})$
 LC: $Y = \exp(a+b \cdot \ln(\text{percent cover}))$
 LD: $Y = \exp(a+b \cdot \ln(\text{basal stem diameter, mm}))$

*** total above ground biomass

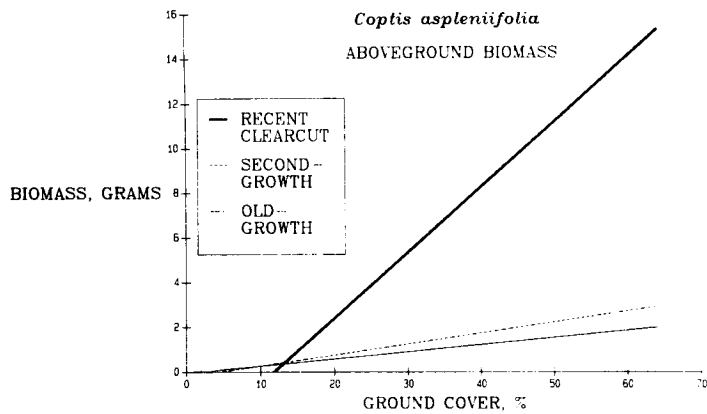


Figure 1. Biomass equations for Coptis asplenifolia in relation to stand age.

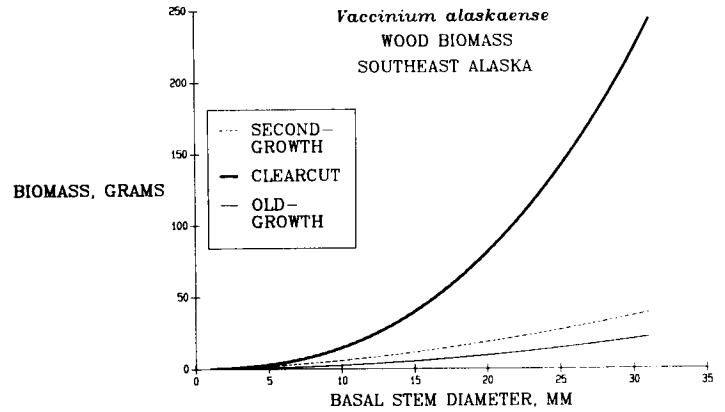


Figure 4. Wood biomass equations for Vaccinium alaskaense in relation to stand age.

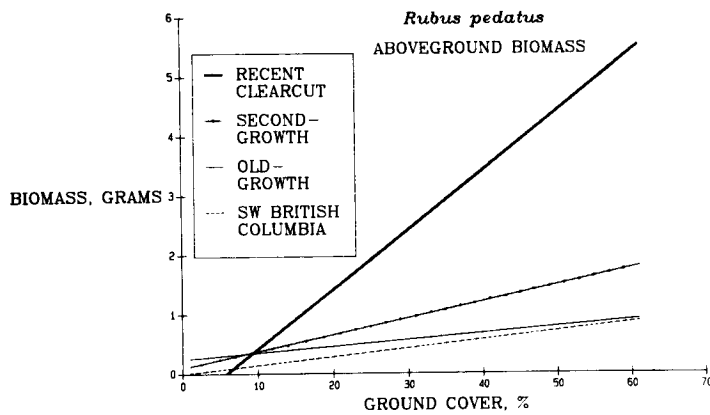


Figure 2. Biomass equations for Rubus pedatus in relation to stand age.

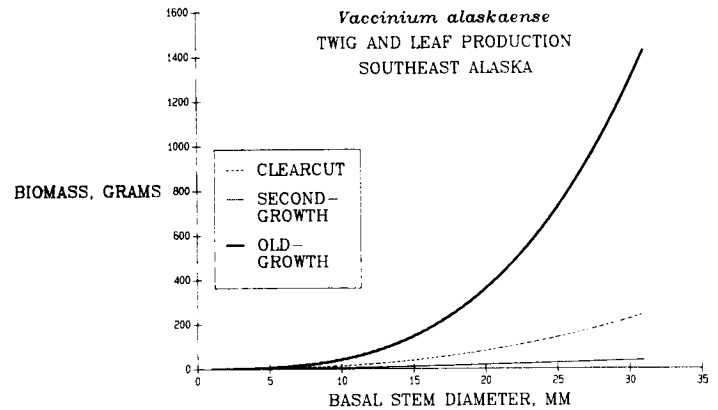


Figure 5. Twig and leaf production equations for Vaccinium alaskaense in relation to stand age.

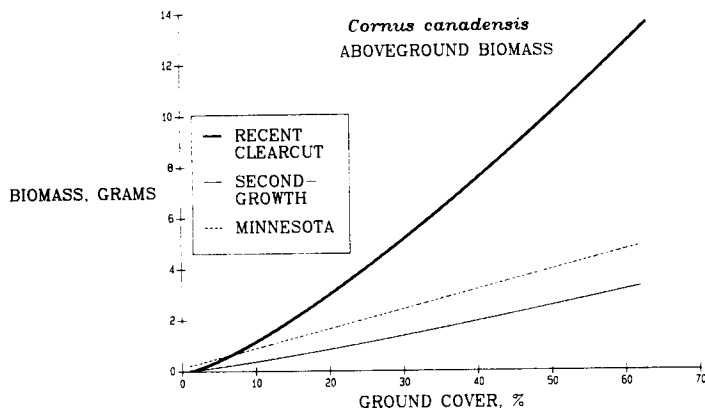


Figure 3. Biomass equations for Cornus canadensis in relation to stand age.

Effects of thinning

Results from experimental manipulations of overstory canopy density generally agreed with results from site age-class comparisons. Thinning resulted in increased vigor of Rubus spectabilis reflected in the equations for woody biomass per unit stem diameter (Table 1, Figure 6). Twig and foliage production decreased per unit stem diameter in response to thinning (Figure 7). The pattern, therefore, generally agreed with that observed for biomass equations from other shrub species across the gradient of overstory canopy openness and shrub density. The largest difference was between the unthinned (control) and the thinned plots equations.

Table 2. Comparisons of biomass equations using F statistics from analyses of covariance for southeast Alaska sites.

Comparison	Species	F statistics***	
		slopes	intercepts
Clearcut vs second growth vs old growth:			
	<u>C. aspleniifolia</u>	** 7.304	-0.950
	<u>C. canadensis</u>	* 4.780	-1.655
	<u>O. horrida</u>	* 4.530	** -8.220
	<u>R. pedatus</u>	* 4.030	0.655
	<u>V. alaskaense</u>	** 11.550	** 23.160
Second growth vs old growth:			
	<u>C. aspleniifolia</u>	* -4.770	** -16.730
	<u>O. horrida</u>	2.040	** -18.700
	<u>R. pedatus</u>	** -9.990	1.170
	<u>V. alaskaense</u>	** -12.770	** -55.980
Clearcut vs second growth:			
	<u>A. felix-femina</u>	2.110	2.620
	<u>C. aspleniifolia</u>	2.330	* 5.530
	<u>C. canadensis</u>	* 8.010	** 92.360
	<u>D. austriaca</u>	** 119.100	** 121.370
	<u>O. horrida</u>	** -18.416	** -20.819
	<u>V. alaskaense</u>	** 58.366	** -82.153
Control vs light thinning vs heavy thinning:			
	<u>R. spectabilis</u>	** -32.500	** -13.800
Control vs light thinning:			
	<u>R. spectabilis</u>	** -37.100	** -39.100
Light vs heavy thinning:			
	<u>R. spectabilis</u>	** -58.600	** -30.700

* significant at $P < 0.95$

** significant at $P < 0.99$

*** F values significantly greater than the critical value indicate in column 1 that slopes are significantly different or in column 2 intercepts are significantly different.

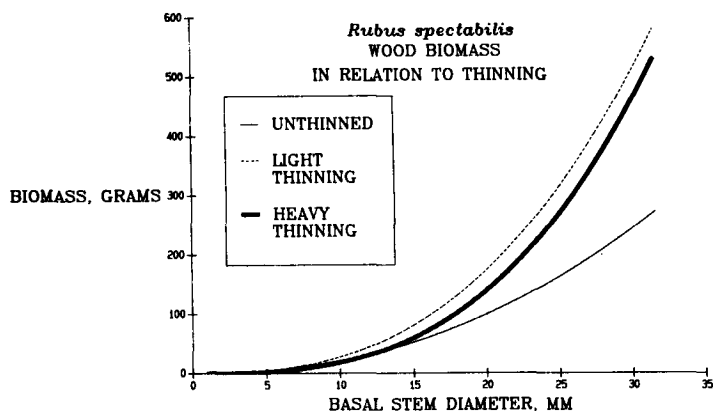


Figure 6. Woody biomass equations for Rubus spectabilis in relation to thinning.

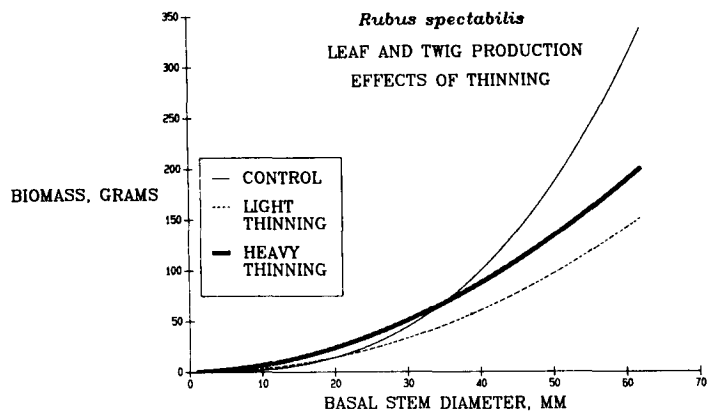


Figure 7. Leaf and twig production equations for Rubus spectabilis in relation to thinning

If woody biomass accumulation per unit stem diameter were proportional to shrub density and light intensity, the heavy thinning rather than the light thinning would be expected to yield the highest biomass estimates. The fact that both light and heavy thinning produced generally similar equations compared with the control (Figure 7) suggests that the effect of thinning on biomass dimension relations may not operate over the restricted range of light and shrub-density values tested, or that other site factors may be involved. For example, although the heavy thinning resulted in 14 times the biomass of the control, and the light thinning resulted in only 1.5 times the biomass of the control for a stem of 25 mm, R. spectabilis growing on the lightly thinned site had the highest proportion of woody biomass per unit stem diameter (Alaback 1980). On the lightly thinned site the second most abundant species V. alaskaense was 17 times more productive than on the heavily thinned site, perhaps offering less competition for canopy development of R. spectabilis because of its more open shade-tolerant canopy than the dense shade-intolerant canopy of R. spectabilis in the heavily thinned site (Alaback 1980). Thus, the balance between woody and canopy components of biomass in shrubs could be a function of light, soils, and competitive interactions with other species.

Regional comparisons

In contrast to many overstory species, most forest understory species are relatively restricted in geographic range so that comparisons between regions are usually difficult to make. Biomass equations were found for some of the species studied in low-elevation Tsuga

heterophylla forests in Alaska in Tsuga mertensiana forests in southwest British Columbia (Yarie 1980), in Picea sitchensis-Tsuga heterophylla forests in coastal Oregon and Pseudotsuga menziesii forests in the Western Cascades (Gholz et al. 1979) mixed conifer forests in northern Idaho and western Montana (Brown 1976), subalpine conifer forests in Glacier National Park, Montana (Gholz et al. 1979) and mixed hardwood-conifer forests in northern Minnesota (Ohmann et al. 1981). Although original data from these other studies were not available for direct statistical comparisons, most biomass estimates from these equations had biases at least as large between regions as was found in this study between age-classes within southeast Alaska, and were therefore assumed to be significantly different.

Only in southwest British Columbia and northern Minnesota were biomass equations developed for herbaceous species found in southeast Alaska. The equations for British Columbia were pooled from three habitat types, which were not significantly different ($P < 0.05$) (Yarie 1980). The data from Minnesota were also based on several habitat types. Only the biomass predictions for the fern Gymnocarpium dryopteris were within 10% of the predictions that used the southeast Alaska equations at a cover value of 50% (Table 3). Although the sites in British Columbia were geographically closer and were also growing in a cool, maritime climate, the biomass equations for those areas were no more similar to those in southeast Alaska than the equations from sites in the extreme continental climate of northern Minnesota. In both regions, equations were found that overestimated biomass for some species and underestimated biomass for others, generally with equal magnitude.

Table 3. Comparisons of biomass equation estimates at 50% cover for herbs from three geographic regions.

Species	----- % of estimate from southeast Alaska -----	
	Northern Minnesota*	southwest British Columbia**
<u>Cornus canadensis</u>	+14.8	--
<u>Gymnocarpium dryopteris</u>	--	-1.61
<u>Maianthemum dilatata</u>	-17.6	--
<u>Rubus pedatus</u>	--	-30.1
<u>Streptopus roseus</u>	+268.4	+199.0
<u>Tiarella trifoliata</u>	--	+225.8

*Ohmann et al. (1981)

**Yarie (1980)

Equations for biomass components of woody shrub species from other geographic regions gave more consistent bias compared with equations from southeast Alaska than did equations for the herbs (Figures 8 and 9). Equations from the northern Rocky Mountains underestimated total biomass for both the riparian species *Oplopanax horridum* and the upland forest species *Menziesia ferruginea*. Biomass equations from Oregon and British Columbia predicted double or more the woody biomass of *Rubus spectabilis* and *Vaccinium alaskaense*. Prediction of twig and foliar biomass from Oregon and British Columbia equations generally underestimated biomass in southeast Alaska for *R. spectabilis* and *V. alaskaense*. The only exception was twig and foliar production of *V. alaskaense* in Oregon, which overestimated biomass relative to the equations from southeast Alaska. The increased biomass and vigor of shrubs from lower latitudes should not be too surprising because these forests are more open and generally grow on better drained, more productive sites (Gholz 1982, Grier and Logan 1977, Yarie 1980).

Conclusions

The dimension analysis technique has been widely accepted as the most efficient and precise means for estimating biomass of both trees and understory vegetation, but more attention needs to be focused on the potential errors of its application. Because understory plants are responsive to short-term changes in microclimate and localized site conditions, a much larger effort is needed in expanding existing libraries of biomass equations to include more site-specific equations. Some of these difficulties with site-specific or area-specific equations could be alleviated by using more precise independent variables, and multiple independent variables such as height x cover, canopy volume, or vigor. In places such as southeast Alaska, however, where understory shrubs often form dense thickets and have highly irregular canopy shapes, estimates of these other variables would likely make the regression technique much less efficient and useful. The simplest and most straightforward approach is to develop new equations for habitat types, age classes, or stand structures that will be the subject of intensive nondestructive sampling. Much of the existing data may also benefit from re-analysis, especially if it can be re-aggregated according to logical ecosystem groups or productivity classes. If more precision in biomass estimation equations is developed, more may be learned of the variation in vigor and structure of understory plants in response to environmental gradients, and more accurate and sensitive analyses of variation in understory growth and abundance will be possible. This increased precision will be of valuable in wildlife habitat assessment and process-oriented ecosystem studies.

Literature Cited

- Alaback, P.B. Biomass and production of understory vegetation in seral Sitka spruce-western hemlock forests of southeast Alaska; Ph.D. diss.; Oregon State University, Corvallis, Oregon. 79p.; 1980.
- Alaback, P.B. Dynamics of understory biomass in Sitka spruce-western hemlock forests of southeast Alaska. *Ecology* 63: 1932-1948; 1982.
- Alaback, P.B. Plant succession following logging in the Sitka spruce-western hemlock forests of southeast Alaska: implications for management. Gen. Tech. Rep. PNW-173. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Forest and Range Experiment Station; 1984.

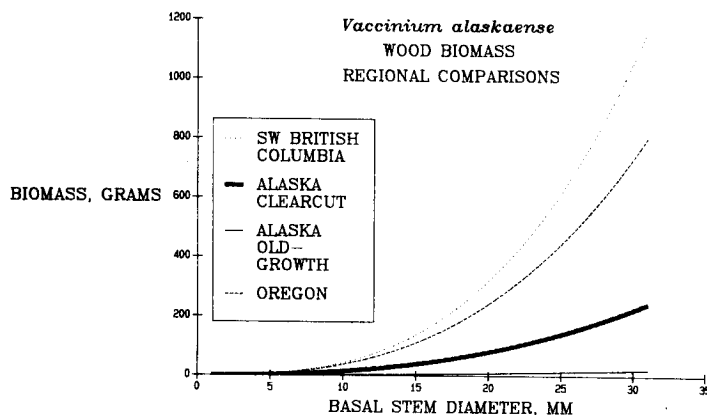


Figure 8. Woody biomass equations for *Vaccinium alaskaense* from different geographic regions.

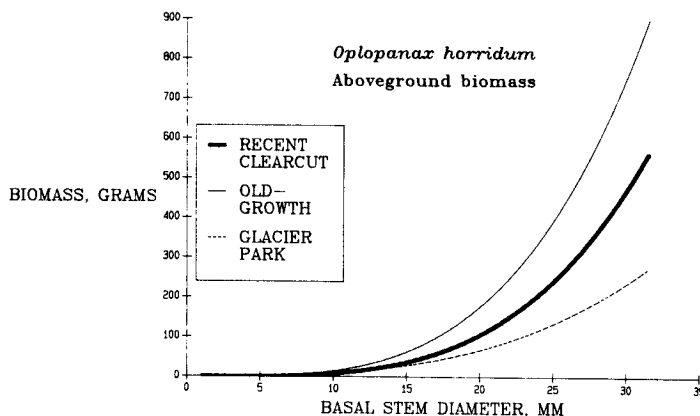


Figure 9. Total aboveground biomass equations for *Oplopanax horridum*

- Alaback, P.B. Biomass regression equations for understory plants in coastal Alaska: effects of species and sampling design on estimates. *Northwest Science* 60:90-103; 1986.
- Baskerville, G.L. Use of logarithmic regression in the estimation of plant biomass. *Can. J. For. Res.* 2:49-53; 1972.
- Brown, J.K. Estimating shrub biomass from basal stem diameters. *Can. J. For. Res.* 6:154-158; 1976.
- Cunia, T. Dummy variables and some of their uses in regression analysis. In: T. Cunia, K. Kuusela, and A.J. Nash, editors, *Proceedings of the June 1973 meeting of IUFRO subject group 54.02. Vol 1.* Nancy, France; SUNY College of Environmental Science and Forestry, Syracuse, New York; 1973.
- Gholz, H.L. Assessing stress in *Rhododendron macrophyllum* through an analysis of leaf physical and chemical characteristics. *Canadian Journal of Botany* 56:546-556; 1978.
- Gholz, H.L. Environmental limits on aboveground net primary production, leaf area, and biomass in vegetation zones of the Pacific Northwest. *Ecology* 63:469-481; 1982.
- Gholz, H.L., C.C. Grier, A.G. Campbell, and A.T. Brown. Equations and their use for estimating biomass and leaf area of plants in the Pacific Northwest. *Res. Pap. 41.*, Corvallis, OR: Forest Research Laboratory, Oregon State University; 1979. 37 p.
- Grier, C.C., and R.S. Logan. Old growth *Pseudotsuga menziesii* communities of a western Oregon watershed: biomass distribution and production budgets. *Ecological Monographs* 47:373-400; 1977.
- Hulten, E. *Flora of Alaska and neighboring territories. A manual of the vascular plants.* Stanford, CA: Stanford University Press; 1968. 1008 p.
- Ohmann, L.F., and D.F. Grigal. Biomass distribution of unmanaged upland forests in Minnesota. *Forest Ecology and Management* 13: 205-222; 1985.
- Ohmann, L.F., D.F. Grigal, and L.L. Rogers. Estimating plant biomass for undergrowth species of northeastern Minnesota. *Gen. Tech. Rep. NC-61.* St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station; 1981. 10p.
- Pastor, J., J.D. Aber, and J.M. Mellilo. Biomass prediction using generalized allometric regressions for some Northeast tree species. *Forest Ecology and Management* 7:265-274; 1984.
- Smith, W.B., and G.J. Brand. Allometric biomass equations for 96 species of herbs, shrubs, and small trees. *Res. Note NC-299.* St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station; 1983. 8 p.
- Stanek, W., and D. State. Equations predicting primary productivity (biomass) of trees, shrubs, and lesser vegetation based on current literature. *Inf. Rep. BC-X-183.* Victoria, BC: Canadian Forest Service Pacific Forest Research Centre; 1978.
- Tappeiner, J.C., II., and H.H. John. Biomass and nutrient content of hazel undergrowth. *Ecology* 54:1342-1348; 1973.
- Whittaker, R.H. Estimation of net primary production of forest and shrub communities. *Ecology* 42:177-180; 1961.
- Yarie, J. The role of understory vegetation in the nutrient cycle of forested ecosystems in the mountain hemlock biogeoclimatic zone. *Ecology* 61:1498-1514; 1980.

BIOMASS ESTIMATES FOR NONTIMBER VEGETATION IN THE TANANA RIVER BASIN OF INTERIOR ALASKA

Bert Mead, John Yarie, and David Herman

Bert R. Mead is a forester and David Herman is a computer programmer, Forest Inventory and Analysis for Alaska Forestry Sciences Laboratory, Pacific Northwest Research Station, USDA Forest Service, 201 East Ninth Avenue, Suite 303, Anchorage, Alaska 99501. John Yarie is an assistant professor of silviculture and forest ecology, Forest Soils Laboratory, University of Alaska, Fairbanks, Alaska 99775-5500.

Vegetation on sample inventory plots in the Tanana River basin was described using horizontal-vertical profile descriptor techniques to show percentage of foliar cover and average height by species. These profiles were used in conjunction with species biomass equations to develop total biomass estimates for each plot. Equations were previously developed for all major species in the area, minor plants were grouped into lifeforms, and the most similar major species equation was applied for weight estimation. Sample weight profiles and a preliminary biomass summary by vegetation type are given. Usefulness of such biomass information for wildlife habitat analysis is discussed.

Introduction

The research unit, Forest Inventory and Analysis for Alaska, USDA Forest Service, has been describing and evaluating all vegetation (timber and nontimber) on inventory plots since 1978. We no longer sample timber to the exclusion of other vegetation resources. This makes good sense for Alaska as much of interior Alaska is currently only marginal in producing wood products; the main economic value of the land may be as fish and wildlife habitat and for oil and mineral production. Identifying areas of good wildlife habitat for a particular animal requires sufficient knowledge of the vegetation to quantify preferred plants in its composition and to identify beneficial patterns of vegetation layering. Monitoring change caused by oil and mineral production requires good baseline vegetation data to adequately assess the impact on the surrounding environment. Also, air pollution originating in industrialized cities in the Soviet Union and Europe is a concern in arctic areas. Plant composition and total biomass changes in various communities may be good indicators of changes created by pollution.

Two of Alaska's first large-area inventories that included nontree vegetation, were done in the 3.6-million-hectare (9-million-acre) Porcupine River basin and the 6.4-million-hectare (16-million-acre) Susitna River basin from 1978 to 1980. Measurement of nontimber vegetation in the Susitna River basin was directed toward estimating annual production of herbage and browse rather than estimating biomass. Annual production was estimated by a double-sampling clip and weigh technique at each sample plot (USDA 1976). Recorded weights were related to foliar cover estimates to project the amount of browse and herbage production for each plot (USDA In press). It became clear that these plot-by-plot weight estimates of annual growth were varying more because of the date of sampling than for any real differences in site productivity.

The problem of obtaining good estimates of productivity is magnified by the short, compressed growing season in Alaska and the elevational differences encountered in large-area river basin inventories. River basin sample units always include extremes of elevation from river bed to surrounding mountainous terrain, whereas conventional timber inventories are generally restricted to the more productive lowlands of Alaska. Attempts were made to adjust the timing of sampling in an already short field season so that plots were measured near the peak of their vegetative development; however, the costs incurred in returning to remote areas to pick up a few isolated high-elevation plots later in the season quickly precluded this as a reasonable solution. The cost of access to much of the sample area also precluded a separate phenological development study whereby subsampled plots would be visited repeatedly to monitor and adjust for observed seasonal differences in phenology. Because of these problems, we concluded that estimates of current annual growth on nontree vegetation were of questionable value. A more appropriate assessment of vegetation seemed to be to describe composition and plant biomass. Data on plant biomass may also be more meaningful for wildlife habitat assessments, which examine total cover and total available browse. Future studies may be able to correlate a site's annual productivity with its biomass and related site factors, such as elevation and aspect.

Biomass estimates have an advantage over browse inventories because the former describe the total resource. No subjective judgments are made on what is available, based on height and twig size considerations, because these vary with yearly snow depth and with differences in use by individual animals. Biomass estimates do not eliminate the sampling problems associated with phenological development but they do minimize them. Estimates are made from percent cover values and applied to equations which were constructed by sampling at sites with maximum seasonal development. Cover estimates on inventory sites which have not reached their full seasonal development are more accurately adjusted than are weight samples of annual growth.

The decision to estimate herbaceous and shrub biomass on our inventory plots was made at the same time we were totally revising our sampling system. We changed from sampling 0.4-hectare (1-acre) homogeneous plots to 8-hectare (20-acre) heterogeneous ground plots, which could include several forest and nonforest vegetation types in each 8-hectare sample. The large plots allowed for collection of more information at a site and were determined to achieve better sampling of edge conditions. It was also more cost effective to sample multiple conditions at fewer sites in a remote area that required helicopter access. We adapted the horizontal-vertical vegetation profile technique developed by the Southeastern Forest Research Station (Cost 1979) as part of the new sample design. This technique, describes vegetation in vertical layers and gives a more complete picture of total browse under a variety of snow conditions; and, it is more descriptive of vegetation in general.

It was against this background that a cooperative study was initiated between FIA and the University of Alaska to develop biomass equations for the Tanana River basin--an area equivalent in size to the States of New York and Connecticut (13.8 million hectares, 34.5 million acres) and with an elevation range from 120 meters (400 feet) at the confluence of the Tanana and Yukon Rivers to 6,096 meters (20,320 feet) at the top of Mount McKinley. See figure 1.

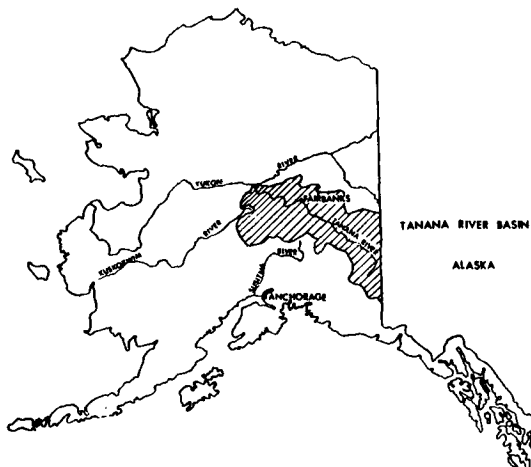


Figure 1. Map of the Alaska showing location of the Tanana River basin.

Methods

Ten of the most common vegetation types in the Tanana River basin were selected to represent the range in vegetation from closed forest to more open shrubland and grassland. Sample locations were selected randomly among the inventory plots that contained these vegetation types. Two transects were located in each sampling location. The first began at a random

starting point. Each transect was 120 meters (400 feet) long; the second line was 50 meters (165 feet) from the first and ran parallel to it. Rectangular biomass sampling plots 0.91 by 0.61 meters (2 by 3 feet) were positioned along the transects every 20 meters (66 feet). A vertical profile was defined by hanging ropes from an aluminum ladder in a technique similar to that described by Harcombe and Marks (1977). The ropes were color coded to delineate 0.3-meters (1 foot) vertical segments.

The ground layer was sampled using a 0.093-square-meter (0.102-square yards) subplot. This sample included all mosses and lichens and any other small plant not exceeding 2.5 centimeters (1.0 inch) in height when fully grown. Once the entire plot was established, the percentage of cover was estimated within each vertical segment starting at 5.3 m (16 feet) and progressing to the ground layer. The material hanging within each segment was clipped. The samples were oven dried and the weight of leaves and twigs determined separately. Frequent checks were made between field crews obtaining biomass information and those working on horizontal-vertical profile plots to determine if significant bias in estimating the percentage of cover was occurring. Results demonstrated repeatability within 10 percent with some of the variation caused by minor differences in vertical layer breaks.

Equations of the form, biomass for 10 cm layer = percent cover multiplied by regression coefficient, were developed from these data by using a conditioned linear regression equation to define the relationship between the percentage of cover and the biomass. Equations were developed for 58 species.

Biomass estimates were made for each inventory plot by using the percentage of cover and height estimates taken on horizontal-vertical profile plots and applying the equations developed for that plant species or by matching its lifeform (as determined by its average size, shape, and weight) to the most similar species for which an equation was available. These data can be summarized in numerous ways: by vegetation types, by natural layers, by arbitrary height classes, or by lifeform grouping.

Discussion

By sampling all the plants in each major vegetation type in the river basin, we were able to develop equations for the most common plants and the plants significant to major wildlife species. Some groups of plants, such as ferns and tall forbs, were missed using this sampling procedure and additional work is needed to develop equations for those lifeforms. Additional study is also needed on shrub biomass of stems and twigs larger than 5 millimeters in diameter. Biomass of these components is highly

variable in relation to the percentage of cover; therefore, they were not included in the initial studies. Estimates of percentage of cover of larger stems and twigs is highly dependent on twig orientation, and biomass of these components may be more accurately assessed by stem size counts.

Tables 1 and 2 show typical plot summaries charted by plant species and by layer. They provide very detailed species-level weight data, which are useful in quantifying occurrence of preferred plant species in wildlife habitats. For example, fruticose and squamulose lichens such as Cetraria and Cladonia are common forage for caribou (Rangifer tarandus) because of the high carbohydrate content and easy digestibility. Other lichens, however, are seldom eaten by caribou. Moose (Alces alces) supplement their browse diets, which consist primarily of willow (Salix sp.) and birch (Betula sp.), with foliose lichens such as Peltigera as well as a variety of other plants. Small mammals, such as snowshoe hare (Lepus americanus), are also known to selectively browse low shrubs and have a strong preference for Vaccinium uliginosum and Ledum groenlandicum. Vaccinium vitis-idaea is an important forage supplement for moose and grizzly bear (Ursus arctos). Although not appearing in the sample

profiles presented, an example of a preferred and less abundant plant in grizzly bear habitat may be Hedysarum whose roots provide essential forage in early spring and late fall. Boykinia is another summer forage plant important to the grizzly bear (United States Department of Interior 1980).

Further summarization of plots within vegetation types is shown in Table 3. Lichens appear to be a common source of forage among all vegetation types. Willow and birch shrubs, which are favored by moose, also represent a substantial portion of the biomass depicted. Preliminary willow and birch biomass estimates for the Tanana River basin are considerably higher than annual browse production figures reported in other studies. This difference is accounted for in older material that is not part of the current year's growth. This second- and possibly third-year material may not be as palatable for moose but is within the twig size reportedly used (up to 4.0 millimeters) (Wolff 1976). Average total biomass estimates from our study appear reasonable when compared with figures of total annual growth reported by Hanley (1984) of 3,266 kilogram/hectare (3,952 pounds/acre) in the understory of an old-growth coniferous forest in Washington.

Table 1. Biomass profile for a single plot in a closed deciduous forest, by plant species and layer.

Species	Ground (vertical extent of layer 0	Layer 2 0 - 1	Layer 3 1 - 4	Layer 4 4 - 20
(kilograms/hectare)				
Mosses:				
<u>Brachythecium</u> spp.	118	--	--	--
<u>Hylocomium splendens</u>	144	--	--	--
<u>Pleurozium schreberi</u>	347	--	--	--
<u>Polytrichum</u> spp.	322	--	--	--
<u>Ptilium crista-castrensis</u>	69	--	--	--
Grasses:				
<u>Calamagrostis canadensis</u>	t ^{a/}	t	133	--
<u>Poa alpigena</u>	t	t	33	--
Shrubs:^{b/}				
<u>Linnaea borealis</u>	30	30	--	--
<u>Ledum groenlandicum</u>	t	t	395	--
<u>Rosa acicularis</u>	t	t	35	--
<u>Vaccinium vitis-idaea</u>	t	461	--	--
Trees:^{b/}				
<u>Alnus crispa</u>	--	--	72	450
<u>Betula papyrifera</u>	--	--	73	--
<u>Picea mariana</u>	--	--	141	1,411

^{a/} Species occurs at trace amounts (<1 kg/ha) with no significant biomass in these layers.

^{b/} Does not include woody material >5 mm in diameter or above 5 meters high.

Table 2. Biomass profile for a single plot in open deciduous forest, by plant species and layer.

Species	Ground	Layer 2	Layer 3	Layer 4	Layer 5
	(vertical extent of layer in decimeters)				
	0	0 - 2	2 - 4	4 - 11	11 - 50
(kilograms/hectare)					
Lichens:					
<u>Cladina</u> spp.	121	--	--	--	--
<u>Cladonia</u> spp.	216	--	--	--	--
Mosses:					
<u>Mnium</u> spp.	26	--	--	--	--
<u>Polytrichum</u> spp.	129	--	--	--	--
Clubmoss:					
<u>Lycopodium annotinum</u>	48	--	--	--	--
Forbs:					
<u>Epilobium angustifolium</u>	t ^{a/}	14	14	--	--
<u>Mertensia paniculata</u>	t	23	--	--	--
<u>Polygonum</u> spp.	t	17	--	--	--
Shrubs: ^{b/}					
<u>Betula occidentalis</u>	t	216	198	1,323	3,510
<u>Betula papyrifera</u>	t	18	37	321	--
<u>Ledum groenlandicum</u>	t	329	691	--	--
<u>Ledum palustre</u>	t	185	83	--	--
<u>Rosa acicularis</u>	t	18	4	--	--
<u>Salix glauca</u>	t	52	35	426	--
<u>Spiraea Beauverdiana</u>	t	1	1	--	--
<u>Vaccinium vitis-idaea</u>	t	146	--	--	--

a/ Species occurs at trace amounts (<1 kg/ha) with no significant biomass in these layers.

b/ Does not include woody material >5 mm in diameter or above 5 meters high.

Table 3. Average biomass by plant grouping and vegetation type for sixteen sample plots.

	Forest	Scrub	Tall Shrub	Low Shrub	Dwarf Shrub	Herbaceous
(kilograms/hectare)						
Lichen	926	993	425	1,167	2,006	572
Moss	1,057	1,524	593	1,197	1,276	870
Grass	164	222	359	283	138	1,209
Forbs	124	60	204	69	62	272
Low shrub	384	333	108	216	234	89
Willow ^{a/} birch ^{a/}	2,396	1,172	4,914	3,390	894	960
Alder ^{a/}	1,682	766	1,899	212	--	309
Tree seedlings ^{a/}	5,892	11,066	1,350	778	--	396
TOTAL	12,625	16,136	9,552	7,312	4,610	4,677
Number of plots	127	23	23	25	7	17

a/ Does not include woody material >5 mm in diameter or above 5 meters high.

Biomass estimates for the river basin provide a good starting point for further wildlife habitat analysis. For instance, willow and birch shrubs occur in most vegetation types; however further characterization shows some of this biomass is above or below browsing reach, of insufficient density, or in areas where terrain or snow depth would prevent utilization. The remaining areas can be evaluated based on knowledge of moose consumption requirements. Studies have shown moose consumption at approximately 5 kilograms per day whereas willow and birch biomass in the Tanana ranges from 100-2,500 kilogram/hectare, depending on the vegetation type. Canadian estimates over extensive areas indicate actual utilization may be in the range of 5-10 percent of the estimated total browse (Telfer 1981). Applying those estimates to the Tanana basin, a productive hectare might provide up to 50 days of browse for moose. However, year-round densities of moose over large areas seldom exceed 12-15 per 1000 hectare (2,500 acres) and are more commonly in the range of 4-6. Because low-protein content limits digestibility of browse, large biomasses may not be food unless additional high-protein forage is also obtainable (Telfer 1981). Our plant species profiles within individual vegetation types will aid in identifying those areas with high-protein forage and with sufficient browse.

The examples given above are not intended to oversimplify the complex task of describing good habitat for a wildlife species. Vegetation descriptions by themselves cannot describe preferred habitats but used in conjunction with terrain descriptions of slope, elevation, and aspect and with snow surveys, they further describe areas likely to provide suitable habitat. Amount of hiding cover, interspersions of forest/nonforest edge, and distance to water; data which are obtained from our photo sampling, will further refine our area estimates of good habitat. This should allow us to assign each plot with wildlife habitat value ratings for several important wildlife species. These value ratings, when tied to vegetation types and other site descriptors, will enable us to produce broad estimates of area of suitable and unsuitable habitat for a particular animal. Further sampling of good habitat areas by interested cooperators would allow them to make estimates of carrying capacity for each wildlife species of interest.

This large data-base of vegetation data could also be applied to the study of habitat requirements of the animals themselves. Very little information is known about the habitat requirements of Alaska's major wildlife species with the possible exception of moose. Analytical study of vegetation and other factors in preferred habitats could identify the important factors in habitat selection and lead to further understanding of the animals themselves.

Conclusions

Biomass and cover estimates provide a useful way to quantify and describe Alaska's vegetation. This presents the new challenge of identifying the vegetation biomass requirements of various wildlife species. These could be expressed as: amount of browse or forage required, amount of preferred plants needed, or as some desirable form of vegetative layering and composition, depending on the wildlife species being studied. This will allow us to make more refined assessments of wildlife habitat than were previously possible. This work has only begun but promises big payoffs in the future.

Literature Cited

- Cost, N. D. Ecological structure of forest vegetation. In: Frayer, W. E., ed. Forest resource inventories, workshop proceedings. 1979 July 23-26; Fort Collins, CO. Fort Collins, CO: Colorado State University; Dept. of Forest and Wood Sciences, 1979: 29-37. Vol. 1 of 2 vol.
- Hanley, T. A. Habitat patches and their selection by wapiti and black-tailed deer in a coastal montane coniferous forest. *Journal of Applied Ecology*. 21:423-436; 1984.
- Harcombe, P. A.; Marks, P. L. Understory structure of a mesic forest in southeast Texas. *Ecology*. 58: 1144-1151; 1977.
- Telfer, Edmund S. Browse inventories: techniques and evaluation. In: Miller, F. L.; Gunn, A., eds. Symposium on census and inventory methods for population and habitats: northwest section, The Wildlife Society: Proceedings. 1980 April 10; Banff, Alberta, Canada. Contribution 217. Moscow, ID: University of Idaho; 1981: 67-82.
- U.S. Department of Agriculture, Soil Conservation Service. Susitna River basin vegetation report. Anchorage, AK: U.S. Department of Agriculture, Soil Conservation Service; [in press].
- U.S. Department of Agriculture, Soil Conservation Service. National range handbook. Washington, DC: U.S. Department of Agriculture, Soil Conservation Service; 1976. 131 p.
- U.S. Department of the Interior, Fish and Wildlife Service. Terrestrial habitat evaluation criteria handbook--Alaska. Anchorage, AK: U.S. Department of the Interior, Fish and Wildlife Service, Division of Ecological Services; 1980. 850 p.
- Wolff, J. O. Utilization of hardwood browse by moose on the Tanana flood plain of interior Alaska. Res. Note PNW-267. Portland, OR: U.S. Department of Agriculture, Forest Service; 1976; 7 p.

RESEARCH PAPERS

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

Moderator: Virgil C. Baldwin, Jr.

"A SUMMARY OF EQUATIONS FOR PREDICTING BIOMASS OF PLANTED SOUTHERN PINES"

V. C. Baldwin, Jr.

Principal Mensurationist, Southern Forest
Experiment Station, Forest Service-USDA,
Pineville, LA 71360

Publications containing regression equations for predicting green or dry weight of whole trees or components of planted southern pine species were reviewed. Their results are summarized with emphasis on the equation databases (sample selection processes, sample size, representiveness), and prediction models utilized. The parameter fitting processes, the measures of precision reported, and the expressed or implied usefulness of the equations are also discussed.

Introduction

Research involving the measurement of whole tree or tree component weight has progressed very rapidly in the South. With some exceptions, green- and dry-weight prediction equations are available for the bole weights of nearly all commercially important softwood and hardwood species. In many instances, branch and foliage weight prediction equations are also available for these species. Baldwin (1984), summarizing statements of earlier researchers, indicated that this work has progressed because of the more widespread use of weight scaling, the utilization of non-bole components in the forest products industry, and the need for biomass as a supplemental energy source.

Forest biomass researchers have responded to meet the need for prediction equations and many have been published. However, because of the great cost involved in developing these equations, there is quite a disparity with respect to which tree components can be weight-predicted. Also, some equations were derived from data sets quite restricted in their representiveness of each species' population and therefore have narrow utility.

The purpose of this paper is to identify, catalog, and describe the data sets, models, and fitted regression equations that have been developed for planted southern pines. This will be an update to, but have a slightly different emphasis than, an earlier review published by Baldwin (1982) and should serve as a guide to the applicability of existing functions for predicting the biomass of southern pine plantations. The published equations and supporting statistics, if reported, will not be given here because of space limitations; however, the references and all essential background information about the publi-

cations are given in Table 1. The reader should go directly to the last page of this table first when referring to it to become familiar with the table-heading codes.

Biomass prediction models

Biomass prediction equations generally have utilized one of the following three basic model forms:

Linear (additive error):

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_j X_j + \epsilon, \quad (1)$$

Non-linear (additive error):

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \dots X_j^{\beta_j} + \epsilon, \quad (2)$$

Non-linear (multiplicative error):

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \dots X_j^{\beta_j} \epsilon, \quad (3)$$

where Y = dependent variables,

X_j = independent or predictor variables,

β_j = parameters to be estimated,

and ϵ = error term.

Model 1 produces multiple linear regressions that can be fitted by standard least squares estimation procedures. Some commonly used variables are diameter at breast height (dbh), diameter squared, total height, diameter squared times total height, age, and live crown ratio.

Model 2 produces nonlinear regression equations that require use of iterative procedures for parameter estimation. Commonly used variables are the same as above.

Model 3 nonlinear regression equations are usually transformed into linear (additive error) regression equations by taking the logarithm of both sides of the equation. In this form the equation parameters can easily be estimated by least squares procedures. However, all common goodness-of-fit statistics will relate to the transformed equation only, and are not directly comparable with the same statistics produced through use of either models 1 or 2 unless further steps are taken. Furthermore, the "intercept" parameter must be corrected to obtain precise prediction when retransforming these equations back into the original units (Baskerville 1972, Yandle and Wiant 1981). Variables commonly used are the logarithmic transformations (to either base e or base 10) of those listed above.

Prediction Equations Available

The earliest green-weight equations for plantation southern pines were developed by McGee (1959) for old-field slash pines in the sandhills of North and South Carolina. The regression equations and any supporting statistics are not given in the paper, but derived tables for bole green weight (including bark) for 4-, 3-, and 2-inch top merchantability limits are presented. McGee's sample consisted of 250 randomly selected merchantable ($\text{dbh} \geq 5$ inches) trees from 150 plantations.

Romancier (1961) produced green-weight and volume tables for plantation-grown loblolly pine for the lower piedmont of Georgia. His sample of 116 trees, representing all crown classes, came from 12 plantations ranging from 19 to 24 years old, 5 to 12 inches dbh, and 30 to 65 feet tall. Equations (from model 1) and tables are for merchantable bole weight or volume predictions only. No goodness-of-fit statistics are given in the paper.

Three other noteworthy studies were done during the early 1960's from which tree component biomass equations were developed (Vaidya 1961, Baker 1962, Rogerson 1964). Foliage-weight equations were developed in all three cases, and in two of the studies branch-weight equations were developed in addition to bole-weight equations. The applicability of these equations is restricted though because the number of sample trees was small, and trees were selected from a single plantation or plantations located in a comparatively small geographic area.

The first of these studies, Vaidya (1961), utilized 12 shortleaf pine trees from 12 plantations and natural stands on the Duke Forest, North Carolina. All trees were dominants or codominants, although they were chosen to represent a broad range of site quality, age, and stand density. The three equations developed from these data, stem dry weight, branch dry weight, and foliage dry weight, were fitted using model 3 form. Standard error of the mean (SE) and coefficient of determination (R^2) statistics are given for each equation.

Baker (1962) sampled 27 dominant and codominant loblolly pine trees from 15 plots on the Duke Forest. They ranged in dbh from 1 to 10 inches, in age from 3 to 31 years, and in site index (base age 50) from 30 to 110 feet. He used model 3 to develop equations to predict stem dry weight, branch dry weight and foliage dry weight as functions of dbh, total height, site index, and percentage of stocking. R^2 and SE statistics are presented for each fitted equation.

Rogerson (1964) fitted loblolly pine foliage data to model 3, with the final equation predicting the weight of loblolly pine foliage as a function of dbh. He also tried projected crown area, tree basal area, and crown length, alone and in combination, in the equation but found that they did not significantly add to the predictions

obtained using dbh alone or dbh and basal area. His 28 sample trees came from one plantation in northern Mississippi. His final equation R^2 was 0.81; SE was not given.

The first large whole-stem dry-weight study published for plantation southern pines was that of Collicott et al. (1968). They sampled 177 slash pine trees from the southwest Georgia area and present merchantable bole, top-volume and dry-weight tables for trees 5 - 12 inches dbh. Their trees were sampled from 54 one-quarter acre plots. The type 1 model form was used. Correlation coefficient (r) and SE values are given for each equation.

The publication by Burkhardt and Clutter (1971) reports various aspects of Burkhardt's biomass research for his doctoral dissertation which involved 702 loblolly pines sampled from 234 old-field plantations in the Georgia Piedmont. Three trees were sampled from each plantation - - 608 of the trees were > 5 inches dbh. Equations and tables are given for green- and dry-weight of the total and merchantable bole with and without bark. Equation form followed the type 1 model and R^2 values were presented. Burkhardt followed this publication with Burkhardt et al. (1972) wherein he presented similar bole-weight equations and tables for plantation loblolly pine sampled from the Virginia Piedmont Region and the Coastal Plains of Virginia, Delaware, Maryland, and North Carolina. They sampled 378 trees from 189 temporary plots which were randomly located within selected unthinned plantations. Equations developed from these data were also type 1 but no goodness-of-fit statistics were given for the weight or volume equations. In both of these studies samples were obtained across as wide a range as practical of tree heights and diameters, plantation ages, sites and densities.

Hasness and Lenhart (1972), Hicks et al. (1972), Hyink et al. (1972), Lenhart (1973), and Lenhart and Hyink (1973) are a series of publications that were based on a data set for plantation loblolly pine that was obtained in east Texas, Arkansas, and Louisiana. Wood only, wood plus bark, and green and dry weight, plus cubic-foot volume tables and prediction equations are given for the tree bole. The equations were based on a sample of 632 trees from 158 unthinned old-field plantations ranging in age from 9 to 30 years, in density from 174 to 1,518 trees per acre, and in site index (base age 25) from 37 to 85 feet. Equations were all from model 1, with both R^2 and SE statistics presented for each equation.

Prediction equations for estimating bole green and dry weight and cubic-foot volume of plantation slash pine in the West Gulf Region were developed by Moehring et al. (1973) and King and Moehring (1974). Their data set was obtained from 379 sample trees felled in 110 plantations established on old fields or previously cutover longleaf (*P. palustris* Mill.) pinelands. Tables are presented with or without a form-class variable for the total and merchantable stem with or without bark. The model 1 form was followed,

$R^2 \geq 0.96$ for all equations, and the addition of a form-class variable in the equations reduced SE in all cases.

Several papers during the 1970's reported research in pine plantations on primary productivity and nutrient content. During this research, trees were felled and weighed, and equations were developed to describe the growth or yield in terms of the various tree components. Although the main thrust of the work was incidental to forest-management-related growth and yield work, the data and equations might prove useful for this purpose under certain circumstances. The following six paragraphs briefly describe these studies:

(1) Swank and Schreuder (1973) sampled 39 white pine (*Pinus strobus* L.) trees from the Coweeta Watershed in western North Carolina over a period of 5 years to determine temporal changes in biomass, surface area, and net production. Equations using a weighted model 1 form to reduce heteroskedasticity were developed to predict foliage, branch, and stem weights at different time periods. Also presented are R^2 and SE statistics.

(2) Annual primary productivity in a loblolly pine plantation was researched by Ralston (1973). He sampled 26 trees in 3 annual collections from 1 study plot in the Piedmont of North Carolina. Equations predicting weight of tree boles, branches, needles, and roots are provided; they are based on model 3. Variance and R^2 supporting statistics are given for each equation.

(3) Nemeth (1973) studied dry matter production in young loblolly and slash pine plantations on the Coastal Plain of North Carolina. Tree component prediction equations using model 3 form were developed from data measurements of 41 loblolly and 15 slash pine trees sampled from 28 permanent plots. Data from both species were combined in the fitting process, and a qualitative variable was included to differentiate the species. However, this species-variable was only significant in the regressions predicting main stem bark, bole needles, and dead branches. Variance of the mean and R^2 values are also included.

(4) Biomass production and nitrogen recovery after fertilization of young loblolly pines in central Mississippi were the topics of Baker et al.'s (1974) research project. They sampled a total of 85 trees 3 to 6 years old. Biomass prediction equations, which are presented in Baker's doctoral dissertation (Baker 1971), were developed for trees in each fertilization treatment. He utilized the model 3 form and provided the correlation coefficient as a goodness-of-fit measure.

(5) Wells et al. (1975) felled 16 trees in a 16-year-old thinned loblolly pine plantation on the North Carolina Piedmont. Biomass prediction equations were developed but were not published; they are available from the authors.

(6) Nelson and Switzer (1975) present prediction equations and tables for green and dry weight of loblolly pines, based on the data from other studies conducted at the Mississippi Agricultural and Forestry Experiment Station (e.g. Baker 1971). The data were originally collected for other study objectives. Their publication presents separate equations (using model 3 form) and tables for estimating aboveground components of plantation loblolly pines at ages 3, 5, 6, 10, and 15. In all, 133 trees were used to develop these equations. Included is the goodness-of-fit measure, R^2 .

Weight and volume equations for southeast coastal plain plantations of loblolly pine were developed by Flowers (1978). He sampled 762 trees from plantations 10 to 30 years old having site indices (base age 25) ranging from less than 45 to more than 84 feet. Similarly, green- and dry-weight equations for southeast coastal plain slash pine are presented in Queen and Pienaar (1977). Their sample also covered a wide range of stand ages, sites, and densities. In both of these publications, results are given for total wood and bark and for "merchantable" portions of the bole only. The equations were developed from model 3 for total cubic foot volume or total green or dry weight. Merchantable volume or weights are predicted using a ratio equation that predicts the proportion of the total volume or weight contained in the merchantable portion of the stem. Those equations are also presented in two recent articles describing growth and yield prediction systems for plantation loblolly and slash pine (Clutter et al. (1984) and Bailey et al. (1982), respectively). R^2 and SE were the support statistics used in these publications.

Edwards and McNab (1979, 1981) sampled trees from young even-aged stands of loblolly, longleaf, shortleaf, and slash pines in central Georgia and developed total tree dry-weight equations for each species. The sample trees ranged in age from 2 to 20 years, in groundline diameter from less than 1 to 6 inches, and in height from about 1 to 29 feet. Equations were of model 3 form. There is no specific reference in the publications stating that these trees came from plantations. They are mentioned here because there is little information available about sapling-size shortleaf and longleaf pines in either natural or planted stands. The goodness-of-fit statistics provided are correlation coefficient and variance of the estimate.

The most recent biomass studies have usually reflected the increased interest in total tree utilization. The resulting publications usually report green- and dry-weight equations for all of the aboveground components, but only one major mensurational study predicts pine plantations stump/root system weights as well as aboveground component weights (Gibson et al. 1984).

Madgwick and Kreh (1980) provided equations to estimate the component biomass of Virginia pine (*Pinus virginiana* Mill.) trees and stands. They fitted data from 506 trees (182 from plantations)

felled from 13 plots in 10 stands on the Virginia Piedmont. The sample included five open-grown trees. The equations developed were from model 3. Also presented are SE values.

Rockwood et al. (1980) produced equations to predict the green and dry weight of dense plantations of Choctawhatchee sand pine (*P. clausa* var. *immuginata* Ward) in the Florida sandhills. Only equations for prediction of green and dry stem weight are given, although all the components were weighed separately; presumably, total tree and/or other component equations could be fitted from the data if needed. The authors anticipate that the main utilization of the species will be for fuelwood. The equation form used was from model 1; R^2 is the only support statistic given. These samples consisted of 28 trees from 3 plantations on the Chipola Experimental Forest.

To study the impact of attacks by the Nantuckett pine tip moth on young loblolly pine, Hedden et al. (1981) looked closely at the effects the defoliation had on the biomass of the tree components. Twenty trees, 10 each from untreated and insecticide-treated plots were felled after a sufficient time of insect attack, and component biomass equations were developed for each 10-tree data set. Model 3 form was used for developing equations. The authors also provided R^2 and SE statistics.

Bower and Clason (1981) reported results of a biomass study utilizing managed plantation loblolly pine in the West Gulf Region. This was a cooperative effort between Louisiana State University and the Weyerhaeuser Company. Component biomass (including root system) models were developed from combined data sets supplied by the two cooperators. The generalized results are reported in the publication, although no prediction equations are provided. The study is one of the first weight-yield studies done in thinned southern pine plantations.

Green-weight, dry-weight, and cubic-foot volume for the aboveground components of unthinned longleaf pines growing in the West Gulf Region can be predicted utilizing equations developed by Baldwin and Saucier (1982, 1983). They sampled 111 trees 1 to 21 inches dbh from 10 plantations in Louisiana and east Texas. Site index (base age 50) ranged from 60 to 119, plantation age from 10 to 44 years, and planting densities from 250 to 1,815 trees per acre. The data were fitted to the model 3 form. Stem weights and volumes to desired top diameters are obtained by predicting weight or volume ratios and then obtaining the proportionate weight from the total weight estimate. Goodness-of-fit statistics included R^2 and SE in logarithmic units, fit index, SE in actual units, coefficient of variation, and percent error.

The most intensive study of crown component biomass yet undertaken in southern pine plantations is reported in Hepp and Brister (1982). They measured and weighed crown components of 364 trees from 182 plots in site-prepared loblolly pine plantations distributed throughout the lower

Coastal Plain of the Carolinas. Total crown, branch, and foliage dry-weight equations are given in the publication. They used model 3 for equation development and included crown ratios, age, and basal area in them. SE and R^2 statistics are also included with the equations.

Shelton et al. (1984) developed equations to predict the weight and volume of plantation-grown loblolly pine trees in the interior flatwoods of Mississippi. They sampled 104 trees ranging in age from 6 to 20 years. Regression equations from the various tree components were developed using models 1 and 2. Ratio equations for estimating the merchantable portion of the main stems are also provided. In addition, they provide support statistics of fit index, SE, and coefficient of variation.

Biomass and nutrient content of a 41-year-old, twice-thinned loblolly pine plantation on a poor site on the upper Piedmont of South Carolina were recently studied by Van Lear et al. (1984). Sixteen trees were felled and regression equations, using model 3 form, were developed from the measurement data for aboveground component biomass. The sample included representative trees from each of the four crown classes found in the plantation. SE and R^2 statistics are also provided.

Gibson et al. (1984) presented preliminary results of an interesting study in which they compared the biomass of planted loblolly, longleaf, shortleaf, and slash pine trees in unthinned plantations on wet, intermediate, and dry sites in northern Louisiana. The stands were long-term study areas in which all four species had been maintained and periodically measured since planting. The stands were 25 years old at the time of sampling. Thirty-six trees of each of the four species were sampled--12 trees on each of the 3 sites. Regression equations (model 3) with R^2 and SE values to predict total-tree green and dry biomass are presented for each species at each location.

Bailey et al. (1985) and Pienaar and Grider (1984) present model 3 equations to predict green weight with bark and dry weight without bark of site-prepared loblolly pine tree boles. Ratio equations are also provided to predict the corresponding merchantable bole weights. R^2 and SE statistics are provided for each equation. These equations were developed from a 472-tree sample taken from 157 well distributed plots on the Piedmont and upper Coastal Plain of Alabama, Georgia, and South Carolina. The trees ranged in size from 1 to 12 inches in dbh, 15 to 70 feet in height, and 40 to 80 feet in site index (base age 25).

Direct-seeded and planted trees were sampled in Lohrey's (1985) study of component biomass of slash pine trees. He sampled 468 trees (201 planted and 267 direct-seeded) from stands throughout central Louisiana. The trees ranged in age from 12 to 48 years. Many of the planted trees were from thinned plantations. Lohrey pro-

vides model 3 green- and dry-weight prediction for all aboveground tree components and ratio equations for prediction of green or dry merchantable stem weight. Equations are presented for direct-seeded trees and plantation trees. The following goodness-of-fit statistics are provided for each equation: fit index, SE in actual units, and coefficient of variation in actual units.

McNab et al. (1985) felled 83 Choctawhatchee sand pine trees in 8 northwestern Florida plantations. They developed model 3 equations to predict the green- and dry-weight and cubic-foot volume of the aboveground components. Their site index range was quite narrow--49 to 53 feet (base age 25)--but the tree ages ranged from 7 to 27 years, dbh from 2 to 12 inches, and total height from 11 to 63 feet. They also used ratio equations to predict merchantable weights or volumes of the bole. The support statistics provided are R^2 and SE.

Baldwin (1986) sampled 130 trees from 12 loblolly pine plantations or study areas in central Louisiana to provide green- and dry-weight prediction equations (model 3) for all aboveground tree components. The data indicated a difference in bole form or taper between trees from unthinned stands and trees from thinned stands, so separate weight-ratio equations are given for these two classes of trees. The trees ranged in age from 9 to 55 years, in dbh from 2 to 21 inches, and in total height from 18 to 94 feet. The goodness-of-fit statistics given are fit index, SE in actual units, coefficient of variation in actual units, and mean percent error.

This author is aware of only three other completed studies that will produce biomass prediction equations within the next year or two for publication. These studies were all accomplished within the Utilization of Southern Timber Research Work Unit of the Southeastern Forest Experiment Station, Athens, Georgia.^{1/} Equations will be provided to predict aboveground components of trees from unthinned slash pine plantations on the coastal plain of south Georgia and north Florida; of trees from unthinned loblolly pine plantations on the Piedmont region of South Carolina, Georgia, and Alabama; and of aboveground and belowground (taproot) components of loblolly pines from northwest Georgia and slash pines from northwest Florida. These will supplement the work of Queen and Pienaar (1977) and Pienaar and Grider (1984) by providing predictions for the non-bole components of trees from those regions.

Sampling Discussion

Henceforth, our discussion will only be concerned with those studies that the author perceives were designed to develop biomass equations for general mensurational use (e.g., growth and yield estimation, stand inventories) rather than for uses incidental to that objective. That leaves us with two groups of papers based on

somewhat arbitrary groupings of the work done in the Southern United States:

(1) The Southeastern States Group - (Virginia, North and South Carolina, Georgia, Alabama, and Florida)

Burkhart et al. (1972),
Madgwick and Kreh (1980),
McGee (1959),
Romancier (1961),
Collicott et al. (1968),
Burkhart and Clutter (1971),
Flowers (1978),
Queen and Pienaar (1977),
Edwards and McNab (1979, 1981),
Rockwood et al. (1980),
Hepp and Brister (1982),
Pienaar and Grider (1984),
McNab et al. (1985).

(2) The Mid-South States Group - Mississippi, Louisiana, Arkansas, and Texas)

Nelson and Switzer (1975),
Shelton et al. (1984),
Hicks et al. (1972)
Hyink et al. (1972),
Moehring et al. (1973),
King and Moehring (1974),
Bower and Clasen (1981),
Baldwin and Saucier (1983),
Lohrey (1985),
Baldwin (1986).

Within the Southeastern States Group, the large growth and yield weight/volume studies tended to have broadly based sampling plans wherein a few trees were taken from many plantations over a wide area within a particular geographic region (e.g., upper or lower Coastal Plain for only the most valuable tree component, the bole, and in some cases only trees ≥ 4.6 inches ee dbh were sampled. The emphasis was on unthinned plantations of loblolly and slash pine on either old-field or mechanically prepared cutover sites. Within these categories the work is relatively complete.

However, there are some notable deficiencies, most of which can only be corrected by more research. Except for the work of Hepp and Brister (1982) and Madgwick and Kreh (1980), branch and foliage components have been largely ignored until recently. The three Southeastern Forest Experiment Station studies mentioned earlier^{2/} will provide prediction equations for aboveground and belowground components of slash and loblolly pines to help in this area. There is also a noticeable absence of data from thinned stands and from older stands. Apparently, there are no equations to predict weights of mature planted longleaf pines in the Southeast.

Within the Mid-South States Group, the larger growth and yield weight/volume studies roughly fall into two subgroups--the Texas studies and the Louisiana-Mississippi studies. The former subgroup was most like the Southeastern States studies, e.g., broadly based, a few trees sampled from many plantations, only bole equations developed for trees ≥ 4.6 inches dbh, and mostly old-

^{1/} Personal communication, Alexander Clark III
USDA-Forest Service

^{2/} Ibid.

field plantations. However, the latter subgroup emphasized tree rather than stand diversity, and developed equations for aboveground tree components. For example, in the Louisiana studies trees were selected across a range of diameter, height, and crown ratio classes within age and site location classes, within thinned and unthinned plantations, and within unthinned and precommercially thinned direct-seeded stands.

Thus, in the Mid-South States Group of studies, indepth complete aboveground tree component weight equations have been developed for loblolly, longleaf, and slash pines over somewhat limited geographical areas within Louisiana and Mississippi. Tree bole weight equations are available for loblolly and slash pines sampled from mostly old-field plantations in eastern Texas, western Louisiana, and southern Arkansas.

A problem with the limited area studies is that the extent of their overall applicability in the Mid-South States is unknown quantitatively. There is no statistical basis to support use of regression equations developed from these studies outside of the plantations from which their data were collected. Indeed this problem is inherent to some degree in all of the studies reviewed. Therefore, all of these equations should be validated for accurate use outside of the populations sampled, and this is true for the entire South. With the exception of the paper by Phillips and Saucier (1979), there is apparently no published information with respect to validation with independent data sets of any of the equations discussed in this paper.

In addition to validation, the possible development of southwide equations for the major pine species should be undertaken, and regional as well as southwide equations still need to be developed for shortleaf pine.

Modeling Discussion

All three basic model forms have been employed by southern researchers in the development of the biomass equations presented. The relative merits of one model over another have been presented elsewhere (e.g., Crow and Laidly 1980, Payandeh 1981, Schlaegel 1982) and are not discussed here. One major problem paramount in all but the most recent studies, however, has been the negligence of researchers in calculating and reporting statistics much more useful than the coefficient of determination (R^2) or the standard error of the estimate ($S_{y,x}$) in order to evaluate goodness-of-fit. Reporting of just a few more useful statistics would also enable readers to make quantitatively based comparisons between published equations. The paper by Schlaegel (1982) addressed this topic. He suggested publication of the following statistics along with all equations:

- (1) Fit index (FI) - a statistic analogous to R^2 (which is identical to R^2 for untransformed equations).

- (2) Standard error of the estimate (S_e) - for transformed equations this statistic needs to be calculated in original units of measure and presented in the untransformed units.
- (3) Coefficient of variation (CV) - based on the untransformed S_e .
- (4) Furnival's Index (I) - a statistic similar in form and information to S_e except that transformed and untransformed equations can be compared on a common basis (Furnival 1961).
- (5) Percent standard error (S(%)) - a goodness-of-fit statistic that describes the mean error of all residuals in a percentage form.
- (6) Percent error (P) - a statistic useful in the comparison of one equation with another.
- (7) The mean ($\bar{x}$) and the sum of squares ($\sum x^2$) - required for a user to be able to construct confidence limits about his predictions obtained from the equation presented.

Another desirable feature of tree component regression equations is that predictions for the components sum to the prediction for the total tree. This is easily accomplished if the same (model 1) form is used for the equations fitted to predict all components (Kozak 1970). However, there are procedures to accomplish this additivity feature even if different model forms are used for different components (Jacobs and Cunia 1980; Cunia and Briggs 1985a, 1985b; Chiyenda and Kozak 1984). Researchers in the South have not indicated their use of these techniques in any of the publications presented in this paper. Additivity of bole components has been achieved in many of the more recent papers through the use of the ratio equation in conjunction with a total bole weight equation to obtain partial bole weights. Generally, a prediction equation for the total tree weight is not fitted; total tree weight is estimated by summing the predictions for the component parts.

Summary and Conclusions

This review provides an update with a slightly different emphasis than an earlier review on this subject by Baldwin (1982). Since that time, about 678 more loblolly, 36 more longleaf, 36 more shortleaf, 504 more slash pine, and 83 more sand pine sample trees have been felled, measured, weighed, and analyzed for the development of tree component biomass equations. Most of the studies were done in the Mid-South States Region as defined earlier, but the largest sample of trees from one region (472 loblolly pines) was obtained from the Southeastern States Region.

Numerous equations have been developed as a result of all this work. Existing model forms

were categorized into three basic types for the purpose of this paper, and the fitted equations published in all the papers were reviewed, classified, and discussed according to the characteristics of each model form. Generally, the authors of these papers have provided little supplemental information with their equations. In many cases (particularly with model 3) a user cannot adequately evaluate an equation's precision or accuracy. Also, there is insufficient information for quantitative equation comparisons.

In general, studies that were broadly based on large areas tended to emphasize sampling of comparatively few trees from many plantations, with the researchers only collecting data sufficient to develop bole weight equations. The studies with less scope in area emphasized sampling of many more trees from fewer plantations, and data were collected for the development of complete tree-component weight equations. For plantations, most of the former work was done in the Southeastern States Region, and the latter work was mostly done in the Mid-South States Region.

The four remaining greatest needs in pine plantation biomass research are (1) development of southwide prediction equations for loblolly and slash pines, (2) development of complete aboveground tree component equations for shortleaf pine, (3) development of more thinned-stand equations in all species, and (4) published validations of new and existing equations.

Literature Cited

- Bailey, R.L.; Grider, G.E.; Rheney, J.W.; Pienaar, L.V. Stand structure and yields for site-prepared loblolly pine plantations in the Piedmont and upper Coastal Plain of Alabama, Georgia, and South Carolina. Research Bulletin 328. Athens, GA: University of Georgia, College of Agriculture Experiment Stations; 1985. 118 p.
- Bailey, R. L.; Pienaar, L. V.; Shiver, B.D.; Rheney, J.W. Stand structure and yield of site-prepared slash pine plantations. Research Bulletin 291. Athens, GA: University of Georgia, College of Agriculture Experiment Stations; 1982. 83 p.
- Baker, J.B. Dry matter production in plantations of loblolly pine. Durham, NC: Duke University, School of Forestry; 1962. 45 p. M.F. thesis.
- Baker, J.B. Response of sapling loblolly pine (*Pinus taeda* L.) to nitrogen fertilization: growth, accumulation and recovery. Mississippi State, MS: Mississippi State University, Department of Agronomy; 1971. 75 p. Ph. D. dissertation.
- Baker, J.B.; Switzer, G.L.; Nelson, L.E. Biomass production and nitrogen recovery after fertilization of young loblolly pines. In: Proceedings of the Soil Science Society of America; 38: 958-961; 1974.
- Baldwin, V.C., Jr. A summary of published southern pine plantation tree biomass studies. In: Proceedings of the 1982 Southern Forest Biomass Working Group Workshop. Alexandria, LA, 1982 June 16-18. New Orleans, LA: U.S. Department of Agriculture, Southern Forest Experiment Station; 1982: 33-45.
- Baldwin, V.C., Jr. The complete tree concept: Its impact on forest mensuration. In: Proceedings of the 1983 Southern Forest Biomass Workshop. Charleston, SC, 1983 June 15-17. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1984: 10-15.
- Baldwin, V.C., Jr. Green- and dry-weight equations for aboveground components of planted loblolly pine trees in the West Gulf Region. Manuscript in preparation. 1986.
- Baldwin, V.C., Jr.; Saucier, J.R. Weight and volume prediction of unthinned, planted longleaf pine. In: Proceedings of the 1981 Southern Forest Biomass Workshop. Georgetown, SC, 1981 June 11-12. Clemson, SC: Clemson University, Belle W. Baruch Forest Science Institute; 1982: 27-38.
- Baldwin, V.C., Jr.; Saucier, J. R. Aboveground weight and volume of unthinned, planted longleaf pine on West Gulf forest sites. Research Paper SO-191. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1983. 25 p.
- Baskerville, G.L. Use of logarithmic regression in the estimation of plant biomass. Canadian Journal of Forest Research. 2: 49-53; 1972.
- Bower, D.R.; Clason, T.R. Biomass estimation for managed loblolly pine plantations. In: Proceedings, In-Place Resource Inventories: Principles and Practices Workshop. Orono, ME, 1981 August 9-14. Washington, DC: Society of American Foresters; 1981: 937-942.
- Burkhart, H.E.; Clutter, J.L. Green and dry weight yields for old field loblolly pine plantations in the Georgia Piedmont. Report No. 22-Series 4. Georgia Forest Research Council; 1971. 11 p.

- Burkhart, H.E.; Parker, R.C.; Strub, M.R.; Oderwald, R.G. Yields of old-field loblolly pine plantations. Publ. FWS-3-72. Blacksburg, VA: Virginia Polytechnic Institute and State University, Division of Forestry and Wildlife Resources; 1972. 51 p.
- Chiyenda, S.S.; Kozak, A. Additivity of component biomass regression equations when the underlying model is linear. Canadian Journal of Forest Research 14:441-446; 1984.
- Clutter, J.L.; Harms, W.R.; Brister, G.H.; Rheney, J.W. Stand structure and yields of site-prepared loblolly pine plantations in the lower coastal plain of the Carolinas, Georgia, and north Florida. General Technical Report SE-27. Asheville, NC: U. S. Department of Agriculture, Forest Service, Southern Forest Experiment Station 1984: 173 p.
- Collicott, L.V.; Strickland, J.R.; Stonecypher, R.W. Dry-weight yields of slash pine plantations--a case study. Technical note No. 27. Bainbridge, GA; International Paper Company, Southlands Experiment Forest; 1968, 17 p.
- Crow, T.R.; Laidly, P.R. Alternative models for estimating woody plant biomass. Canadian Journal of Forest Research 10:367-370; 1980.
- Cunia, T.; Briggs, R.D. Forcing additivity of biomass tables: use of the generalized least squares method. Canadian Journal of Forest Research 15:23-28; 1985a.
- Cunia, T.; Briggs, R.D. Harmonizing biomass tables by generalized least squares. Canadian Journal of Forest Research 15:331-340; 1985b.
- Edwards, M.B.; McNab, W.H. Biomass prediction for young southern pines. Journal of Forestry 77:291-292. 1979.
- Edwards, M.B.; McNab, W.H. Biomass prediction for young southern pines--An Addendum. Journal of Forestry 79:291. 1981.
- Flowers, W.R., Jr. Individual tree weight and volume equations for site prepared loblolly pine plantations in the coastal plain of the Carolinas, Georgia, and north Florida. Athens, GA: University of Georgia; 1978. 53 p. M.S. Thesis
- Furnival, G.M. An index for comparing equations used in constructing volume tables. Forest Science. 7:337-341; 1961.
- Gibson, M.D.; McMillin, C.W.; Shoulders, E. Preliminary results for weight and volume of even-aged, unthinned, planted southern pines on three sites in Louisiana. In: Saucier, J.R., ed., Proceedings of the 1984 Southern Forest Biomass Workshop, Athens, GA, 1984 June 5-7. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1984: 69-79.
- Hasness, J.R.; Lenhart, J.O. Cubic foot volumes for loblolly pine trees in old-field plantations in the interior West Gulf Coastal Plain. Texas Forest Paper No. 12. Nacogdoches, TX: Stephen F. Austin State University, School of Forestry; 1972. 7 p.
- Hedden, R.L.; Richmond, J.A.; Thomas, H.A.; Lashomb, J.H. Impact of Nantucket pine tip moth attack on young loblolly pine biomass. In: Barnett, J.P., ed, Proceedings of the First Biennial Southern Silviculture Research Conference, 1980 Nov 6-7, Atlanta, GA: New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station, General Technical Report SO-34; 1981: 238-241.
- Hepp, T.E.; Brister, G.H. Estimating crown biomass in loblolly pine plantations in the Carolina Flatwoods. Forest Science. 28(1): 115-127; 1982.
- Hicks, D.R.; Lenhart, J.D.; Somberg, S.I. Merchantable green weights for loblolly pine trees in old field plantations in the interior West Gulf Coastal Plain. Texas Forestry Paper 16. Nacogdoches, TX: Stephen F. Austin State University, School of Forestry; 1972. 4 p.
- Hyink, D.M.; Lenhart, J.D.; Somberg, S.I. Oven-dry weights for loblolly pine trees in old-field plantations in the interior West Gulf Coastal Plain. Texas Forestry Paper No. 17. Nacogdoches, TX: Stephen F. Austin State University, School of Forestry; 1972. 4 p.
- Jacobs, M.W.; Cunia, T. Use of dummy variables harmonize tree biomass tables. Canadian Journal of Forest Research. 10:483-490; 1980.
- King, A.L.; Moehring, D.M. Dry-weight tables for slash pine in east Texas. Research Note 2. College Station, TX: Texas A & M University, Department of Forest Science; 1974. 3 p.
- Kozak, A. Methods for ensuring additivity of biomass components by regression analysis. Forestry Chronicle. 46: 402-404; 1970.
- Lenhart, J.D. Green and dry weight yields for unthinned old-field loblolly pine plantations in the interior West Gulf Coastal Plain. Texas Forestry Paper No. 18. Nacogdoches, TX: Stephen F. Austin State University, School of Forestry; 1973. 12 p.
- Lenhart, J.D.; Hyink, D.M. Direct estimation of oven-dry weight of plantation grown loblolly pine trees in the interior West Gulf Coastal Plain. Forest Products Journal. 23(7):49-50; 1973.
- Lohrey, R.E. Aboveground biomass of planted and direct-seeded slash pine in the West Gulf Region. In: Saucier, J.R., ed., Proceedings of the 1984 Southern Forest Biomass Workshop; Athens, GA, 1984 June 5-7. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1985 75-81.

- Madgwick, H.A.I.; Kreh, R.E. Biomass estimation for Virginia pine trees and stands. *Forest Science*. 26(1):107-111; 1980.
- McGee, C.E. Weight of merchantable wood with bark from planted slash pine in the Carolina sandhills. Research Note No. 128, Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Experiment Station; 1959. 2 p.
- McNab, W.H.; Outcalt, K.W.; Brendemuehl, R.H. Weight and volume of plantation-grown Chocotawhatchee sand pine. Research Paper SE-252. Asheville, NC; U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1985. 44 p.
- Moehring, D.M.; Holmes, M.R.; Merrifield, R.G. Tables for estimating cubic foot volume and fresh weight of planted slash pine. Publication MP-1079. College Station, TX: Texas A & M University, Texas Agricultural Experiment Station; 1973. 28 p.
- Nelson, L.E.; Switzer, G.L. Estimating weights of loblolly pine trees and their components in natural stands and plantations in central Mississippi. Technical Bulletin 73. Mississippi State, MS: Mississippi State University, Mississippi Agricultural and Forestry Experiment Station; 1975. 14 p.
- Nemeth, J.C. Dry matter production in young loblolly (*Pinus taeda* L.) and slash pine (*Pinus elliotii* Engelm.) plantations. *Ecological Monographs*. 43(1): 21-41; 1973.
- Payandeh, B. Choosing regression models for biomass prediction equations. *Forestry Chronicle*. 57:229-232; 1981.
- Phillips, D.R.; Saucier, J.R. A test of prediction equations for estimating hardwood understory and total stand biomass. Georgia Forest Research Paper 7. Macon, GA: Georgia Forestry Commission, Research Division; 1979. 8 p.
- Pienaar, L.V.; Grider, G.E. Standard volume and weight equations for site-prepared loblolly pine plantations in the Piedmont of South Carolina, Georgia and Alabama. Technical Report 1984-3. Athens, GA: University of Georgia, School of Forest Resources, Plantation Management Research Cooperative; 1984. 13 p.
- Queen, W.R.; Pienaar, L.V. Green and dry stem weight equations for site-prepared slash pine plantations in the coastal plain of Georgia and north Florida. Research Paper 1. Athens, GA: University of Georgia, School of Forest Resources, Plantation Management Research Cooperative; 1977. 58 p.
- Ralston, C.W. Annual primary productivity in a loblolly pine plantation. In: IUFRO Biomass Studies, Proceedings of the Working Party on the Mensuration of the Forest Biomass, Vancouver, B.C., Canada. 1973 Aug. 20-24, International Division of Forestry Research Organizations. 54.01 Mensuration, Growth and Yield; 1973: 105-117.
- Rockwood, D.L.; Conde, L.F.; Brendemuehl, R.H. Biomass production of closely spaced Chocotawhatchee sand pines. Research Note SE-293. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1980. 6 p.
- Rogerson, T.L. Estimating foliage on loblolly pine. Research Note SO-16. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1964. 3 p.
- Romancier, R.M. Weight and volume of plantation-grown loblolly pine. Research Note 161. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1961. 2 p.
- Schlaegel, B.E. Testing, reporting, and using biomass estimation models. In: Gresham, Charles A., ed., Proceedings of the 1981 Southern Forest Biomass Workshop. Georgetown, SC, 1981 June 11-12. Clemson, SC: Clemson University, Belle W. Baruch Forest Science Institute; 1982: 95-112.
- Shelton, M.G.; Nelson, L.E.; Switzer, G.L. The weight volume and nutrient status of plantation-grown loblolly pine trees in the interior flatwoods of Mississippi. Technical Bulletin 121. Mississippi State, MS: Mississippi State University, Mississippi Agricultural and Forestry Experiment Station; 1984. 27 p.
- Swank, W.T.; Schreuder, H.T. Temporal changes in biomass, surface area, and net production for a *Pinus Strobus* L. forest. In: Proceedings of the Working Party on the Mensuration of the Forest Biomass, Vancouver, B.C., Canada, 1973 August 20-24. IUFRO Biomass Studies, IUFRO, S4.01 Mensuration Growth and Yield; 1973: 171-182.
- Vaidya, M.S.L. Dry matter production and nutrient accumulation in plantations of short leaf-pine. Durham, NC: Duke University, School of Forestry; 1961. 72 p. D.F. dissertation.
- Van Lear, D.H.; Waide, J.B.; Teuke, M.J. Biomass and nutrient content of a 41-year-old loblolly pine (*Pinus taeda* L.) plantation on a poor site in South Carolina. *Forest Science*. 30(2):395-404; 1984.

Wells, C.G.; Jorgensen, J.R.; Burnette, C.E.
Biomass and mineral elements in a thinned
loblolly pine plantation at age 16. Research
Paper SE-126. Asheville, NC: U.S. Department
of Agriculture, Forest Service, Southeastern
Forest Experiment Station; 1975. 10 p.

Yandle, D.O.; Wiant, H.V. Biomass estimates based
on the simple power function-some calcula-
tion considerations. In: Proceedings, In-Place
Resource Inventories: Principles and Practices
Workshop. Orono, ME, 1981 August 9-14.
Washington, D.C.: Society of American
Foresters; 1981: 350-354.

Table 1.--A summary of published biomass data, functions, and tables, by species, for southern pine plantation trees

Reference	Location State	Sample size: No.	Age Yr	Data range				Tree component ^{1/}								Weight ^{2/}		Prediction ^{3/} equation model
				Dbh In	Height Ft.	Site index Ft.	Density Trees/acre	S/R:	TW	TS	MS	BR	F	B	Green	Dry		
LOBLOLLY PINE																		
Romancier 1961	GA Piedmont	116	19-24	5-12	30-65	-	-							x			x	1
Baker 1962	NC Piedmont	27	3-31	1-10	5-69	30-110 (base age 50)	430-1,816 (at cut)	x	x					x	x		x	3
Rogerson 1964	MS	28	25	5-11	37-66	69-97 (base age 50)	-								x		x	3
Baker 1971 Baker et al. 1974	Central MS	85	3-6	0-3	2-19	80 (base age 50)	910 (planted)	x	x					x	x	x	x	3
Burkhart & Clutter 1971	GA Piedmont	702 total 608 >5 in dbh	9-32	3-12	15-80	25-80 (base age 25)	300-1,675 (at cut)			x	x				x		x	1
Burkhart et al. 1972	VA Piedmont & Coastal Plain of VA,DE, MD, NC	378	9-35	-	-	47-84	300-2,900			x	x				x		x	1
Hicks et al.1972 Hyink et al.1972 Lenhart& Hyink 1973	Interior coastal plain of east TX, ARK, LA	632	9-30	5-19 (99% were 5-13)	25-85 (97% were 30-70)	37-85 (base age 25)	174-1,518			x	x				x		x	1
Nemeth 1973	NC Coastal Plain	41	4-11	-	-	-	546 (at cut)	x	x	x	x	x	x	x			x	3
Ralston 1973	NC Piedmont	26	13-15	-	-	-	681(planted) 606 (at cut)	x	x	x	x	x	x				x	3

See footnotes at end of table.

Table 1.--A summary of published biomass data, functions, and tables, by species, for southern pine plantation trees (continued)

Reference	Location	Sample size	Age	Data range				Tree component ^{1/}								Weight ^{2/}		Prediction ^{3/} equation model
				Dbh	Height	Site index	Density	S/R:	TW	TS	MS	BR	F	B	Green	Dry		
<u>LOBLOLLY PINE</u>																		
Nelson & Switzer 1975	Central MS	133	3-15	1-11	4-56	65-105 (base age 50)	-	x	x		x	x	x	x	x		3	
Wells et al. 1975	NC	16	16	6.5 (avg.)	49 (avg.)	68 (base age 25)	907 (at cut)	(equations not published; available from authors)										
Flowers 1978 Clutter et al.1984	Coastal Plain North FL	762	10-30	2-14	15-75	<45 to 84>			x	x			x	x	x		2,3	
Edwards & McNab 1979, 1980	GA Piedmont	25	2-8	1-4 (4/)	2-19	-	-	x							x		3	
Hedden et al.1981	MD	20	10	4-10	14-27		4,840 (planted)	x	x		x	x			x	x	3	
Gibson et al.1984	LA	36	25	2-9	23-66	-	1,210 (planted)	x							x	x	3	
Pienaar & Grider 1984	AL	472	10-25	1-12	15-70	40-80 (base age 25)	300-1,500								x	x	2,3	
Bailey et al.1985	GA SC								x	x								
Shelton et al.1984	MS	104	6-20	1-14	8-76	70-82 (base age 25)	486-1,740	x	x	x	x	x	x	x	x	x	1,2	

See footnotes at end of table.

Table 1.--A summary of published biomass data, functions, and tables, by species, for southern pine plantation trees (continued)

Reference	Location	Sample size	Age	Data range				Tree component ^{1/}							Weight ^{2/}		Prediction ^{3/} equation model
				Dbh	Height	Site index	Density	S/R	TW	TS	MS	BR	F	B	Green	Dry	
	State	No.	Yr	In	Ft	Ft	Trees/acre										
<u>LOBLOLLY PINE</u>																	
VanLear et al. 1984	SC	16	41	5-15	76 (avg.)	-	177	x		x	x	x	x		x		3
Baldwin 1986	LA	130	9-55	2-21	18-94	40-76 (base age 25)	13-1,133			x	x	x	x	x	x	x	2,3
<u>SLASH PINE</u>																	
McGee	NC & SC	250	9-25	5-12	20-75	-	-				x				x		(only tables published)
Collicott et al. 1968	GA	177	17	5-12	-	-	-			x	x				x		1
Moehring et al. 1973	East TX	379	-	5-14	30-85	-	-			x	x				x		1
Nemeth 1973	NC Coastal Plain	15	4-11	-	-	-	545 (at cut)	x	x	x	x	x	x	x	x		3
King & Moehring 1974	East TX	220	-	5-23	27-87	-	-			x	x				x		1
Queen & Pienaar 1977	GA	686	11-28	3-12	20-75	41-81 (base age 25)	250-650			x	x				x	x	2,3
Bailey et al 1982	FL																
Edwards & McNab 1979, 1981	Central GA	29	3-8	1-6	2-19	-	-	x							x		3
Gibson et al. 1984	LA	36	25	2-9	23-66	-	1,210	x							x	x	3
Lohrey 1985	LA	468	12-48	1-19	12-110	30-78 (base age 25)	51-3,580			x	x	x	x	x	x	x	2,3

See footnotes at end of table.

Table 1.--A summary of published biomass data, functions, and tables, by species, for southern pine plantation trees (continued)

	:	:	:	:	Data range				:	Tree component ^{1/}								:	Weight ^{2/}	:	Prediction ^{3/}						
Reference	:	Location	:Sample Size:	Age	:	Dbh	:Height:	Site index	:	Density	:	S/R:	TW	:	TS	:	MS	:	BR	:	F	:	B	:	Green:	Dry:	equation model
		<u>State</u>	<u>No.</u>	<u>Yr</u>		<u>In</u>	<u>Ft</u>	<u>Ft</u>		<u>Trees/acre</u>																	
<u>LONGLEAF PINE</u>																											
Edwards & McNab 1979, 1980		Central GA	24	2-16		1-4	1-29	-		-			x												x		3
Baldwin & Saucier 1982, 1983		Central LA & East TX	111	10-44		1-21	12-83	60-119 (base age 50)		250-1,815		x	x	x	x	x	x	x	x	x	x	x	x	x	x		2,3
Gibson et al. 1984		LA	36	25		2-9	23-66			1,210		x												x	x		3
<u>SHORTLEAF PINE</u>																											
Vaidya 1961		NC	12	-		1-9	8-66	-		-				x	x	x									x		3
Edwards & McNab 1979		Central GA	22	4-20		1-4	3-17	-		-		x													x		3
Gibson et al. 1984		LA	36	25		2-9	23-16	-		1,210		x												x	x		3
<u>WHITE PINE</u>																											
Swank & Schreuder 1973		Western NC	39	10-15		1-9	11-45	-		1,760-1,790 (at cut)				x			x	x							x		1

See footnotes at end of table.

Table 1.--A summary of published biomass data, functions, and tables, by species, for southern pine plantation trees (continued)

Reference	Location	Sample size	Age	Data range				Tree component ^{1/}										Weight ^{2/}	Prediction ^{3/}
				Dbh	Height	Site index	Density	S/R	TW	TS	MS	BR	F	B	Green	Dry	equation model		
	State	No.	Yr	In	Ft	Ft	Trees/acre												
<u>SAND PINE</u>																			
Rockwood et al. 1980	FL sandhills	28	6-17	1-5	9-45	-	3,630-5,808 (planted)	x	x	x	x	x	x	x	x	x			1
McNab et al. 1985	FL	83	7-27	2-12	11-63	49-53 (base age 25)	225-955	x	x	x	x	x	x	x	x	x			2,3
<u>VIRGINIA PINE</u>																			
Madgwick & Kreh 1980	VA Piedmont	182	8	2-3	13-16	83 (base age 50)	619 (at cut)	x	x			x	x	x		x			3

^{1/} Code indicating the tree components measured: S/R = all or part of the stump-root system
 TW = total tree weight
 TS = total stem (bole) weight
 MS = merchantable stem weight
 BR = branches
 F = foliage
 B = bark

^{2/} Study emphasized either green or dry weight as indicated.

^{3/} Model 1 is $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_j X_j + \epsilon$,

Model 2 is $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \dots X_j^{\beta_j} + \epsilon$,

Model 3 is $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \dots X_j^{\beta_j} \epsilon$ transformed to $\ln(Y) = \beta_0' + \beta_1 \ln(X_1) + \beta_2 \ln(X_2) + \dots + \beta_j \ln(X_j) + \epsilon'$

^{4/} Ground-line diameter.

SUMMARY OF BIOMASS EQUATIONS AVAILABLE FOR
SOFTWOOD AND HARDWOOD SPECIES IN THE SOUTHERN
UNITED STATES^{1/}

Alexander Clark III

Research Wood Scientist, USDA Forest Service,
Southeastern Forest Experiment Station, Forestry
Sciences Laboratory, Athens, GA, 30602.

Abstract--Published equations for estimating biomass of entire trees and tree components for softwoods and hardwoods > 1.0 inches d.b.h. are summarized by species, geographic location, and dependent and independent variables. References containing 244 sets of equations for 11 softwood and 52 hardwood species are listed. Regression models commonly used to estimate total-tree and component biomass are also reviewed.

Researchers in the South have developed a large number of tree biomass prediction equations for individual species and species groups. Timber owners, foresters, and researchers need these equations to estimate biomass of forest stands for forest management and utilization decisions. To obtain these biomass estimates, they need to know what species equations are available, what their independent and dependent variables are, and in what geographic areas the equations were developed. Several reviews of the biomass literature (Keays 1971, Stanek and State 1978, Madgwick 1976, Art and Marks 1971, Young 1976, Hitchcock and McDonnell 1979, Tritton and Hornbeck 1982, Clark 1982, Phillips 1982, Baldwin 1982, McNab et al. 1982, and USDA Forest Service, 1984) summarize the many studies conducted to develop biomass equations for entire trees and tree components. This paper updates these literature reviews of biomass research in the South and summarizes the biomass equations available for softwood and hardwood species 1.0 inch d.b.h. and larger growing in natural stands. In this review, the South is defined as the area extending from Virginia and West Virginia in the east to eastern Oklahoma and Texas in the west.

Tree Biomass Prediction Equations

Researchers have collected biomass data on thousands of softwood and hardwood trees in the South to develop equations for predicting green and dry weights and volumes of entire trees and their components. The biomass equations developed for southern hardwoods are summarized in table 1 by species and location. Table 2 summarizes the equations developed for softwoods by species and location. These tables also show the number of trees sampled, the independent and dependent variables used in the equations, and the investigator or author.

^{1/}Paper presented at the Tree Biomass Regression Functions and their Contribution to the Error of Forest Inventory Estimates Workshop, Syracuse, NY, May 27-30, 1986.

Table 1 lists 180 sets of hardwood equations developed to estimate the biomass of 52 hardwood species and Table 2 lists 64 sets of softwood equations developed to estimate the biomass of 11 softwood species. These species account for over 92 percent of the volume on commercial forest lands in the South and make up over 65 percent of the commercial species listed by the USDA Forest Service's Renewable Resource Survey Unit in the Mid-South and Southeast.

No southwide hardwood species equations have been developed. However, regional weight equations for hardwoods are available for the Gulf and Atlantic Coastal Plains (Clark et al. 1985b), the Piedmont Plateau (Clark et al. 1986b), the Southern Appalachian Mountains (Clark and Schroeder 1986), and the Upland-South^{2/} (Clark et al. 1986c). Biomass data for hardwood species sampled in Alabama, Georgia, Florida, South Carolina, and western North Carolina have been pooled for the development of southeastern hardwood species equations (Clark et al. 1986a). Only statewide or local sampling has been conducted in east Texas, Louisiana, and Mississippi, and no regional hardwood species equations are available for the Deep South.

Statistical comparisons for southern hardwood species, similar to those reported by Jacobs and Monteith (1981) for species in the Northeast, have not been published. However, to illustrate the differences in total-tree green weight that exist between geographic regions, equations for several important hardwood species were applied to trees with identical d.b.h. and total height (Clark 1982a, 1982b). Results indicate differences of up to 20 percent for southern red oak, white oak, and hickory in comparisons between the Southeast and east Texas and Oklahoma.

The within-species geographic variations in specific gravity (dry weight per cubic foot), moisture content, and green weight of wood and bark per cubic foot of wood have been statistically compared for eight southern hardwood species (Clark et al. 1986d). The species examined include yellow-poplar, sweetgum, red maple, white oak, scarlet oak, hickory, blackgum, and ash. The results indicate that stemwood specific gravity of the species sampled in the Coastal Plains, Piedmont, Southern Appalachian Mountains, and Upland-South differ significantly ($P=0.05$) among geographic regions. However, there was no indication of a consistent geographic trend except for red maple, whose wood specific gravity decreased from north to south.

For each species except for blackgum, wood moisture content varied significantly among geographic regions. The green weight of wood and bark per cubic foot of wood or weight-scaling factor also varied significantly among geographic regions for all species except hickory.

^{2/}Upland-South includes north Alabama, Mississippi, eastern Arkansas, southern Kentucky, and Tennessee.

However, the weight-scaling factors for sweetgum, red maple, and white oak did not differ significantly between the Piedmont and Coastal Plains. Differences in weight-scaling factors among geographic regions within a species were as high as 16 percent. Because of the magnitude of the within-species regional differences in wood specific gravity, moisture content, and green weight per cubic foot, regional, rather than specieswide, green and dry weight prediction equations and weight-scaling factors should be used for marketing hardwood timber. Species-wide equations are suitable for resource surveys of the Southeast or Mid-South.

Numerous equations have been generated for southern softwoods in natural stands. Equations are available for the four major southern pines, white pine, Virginia pine, sand pine, pond pine, eastern hemlock, eastern redcedar, and pond cypress. Equations for the four major southern pine species have been developed for individual states, such as Georgia (Saucier et al. 1981) and east Texas (Walters 1982), and for geographic regions, such as the Piedmont or Coastal Plain of the Southeast (Saucier et al. 1985) or a group of states (Clark 1986). No southwide pine biomass equations have been developed, however.

As with the hardwoods, there have been no published statistical comparisons of results of equations developed for softwood species at different locations. However, graphs of plotted data for trees at different locations but with identical d.b.h. and total heights show smaller differences for softwoods than that reported for hardwoods. The four major southern pines do increase in weight per cubic foot from north to south across their natural ranges (Clark 1982a), however.

McNab and others (1982) compared results from five equations developed for loblolly pine from Texas to Georgia and found only a 3 to 6 percent difference among them. Clark (1982a) compared predictions for slash pine from Georgia, Alabama, and Mississippi and found a maximum difference of 10 percent. The maximum difference among the four major pines was reported for shortleaf pine. For it, samples in northern Mississippi differed by 10 to 17 percent from samples in Georgia, Texas, and Oklahoma (McNab et al. 1982).

For regional resource surveys, the available state or Southeast area equations for major southern pine species appear to be sufficient. However, these equations need to be tested locally before they are used to predict market yields for southern pines.

Regression Models Used to Estimate Biomass

Researchers have used a variety of regression models for estimating total-tree and tree-component biomass in the South. The three most common models are the linear, weighted linear, and log-log. The linear models are the single-variable

$$Y = a + b(D^2H) + \epsilon \quad (1)$$

or the two-variable

$$Y = a + b(D^2) + c(H) + \epsilon \quad (2)$$

where: Y = predicted component weight or volume
D = tree diameter at breast height in inches
H = tree-total height, height to 4-inch d.o.b. top, or saw-log merchantable height in feet
 ϵ = experimental error
a, b, c = regression coefficients.

The variance of total-tree and total-stem weight is heterogeneous. That is, when plotted over D^2H , variance increases with increasing D^2H . A weighted or log-log transformation, therefore, is commonly used to obtain a more homogeneous variance. The form of the weighted linear model is:

$$\frac{Y}{(D^2H)^K} = \frac{a}{(D^2H)^K} + b(D^2H) (D^2H)^K + \epsilon \quad (3)$$

where: K = weighting factor.

The weighting factor (K) is computed using the procedure reported by McClure et al. (1983). The logarithmic transformation is another common method for obtaining a relatively homogeneous variance. The log-log model is:

$$\text{Log } Y = a' + b \text{ Log } (D^2H) + \epsilon' \quad (4)$$

where: $a' = \log a$
 $\epsilon' = \log \epsilon$

When developing biomass equations from hardwood tree data collected on 25 1/10 acre plots in the Coastal Plain, we found a single log-log equation resulted in biased estimates when fitted to trees 1 to 20 inches d.b.h. It overestimated the weight of trees < 10 inches and underestimated the weight of trees > 12 inches. This bias probably occurred because of the large number of observations in the smaller diameter classes caused by the area sampling procedures used to select sample trees. We used a segmented log-log equation to correct for this bias (Clark et al. 1985a). The segmented procedure used two equations--one for trees < 11.0 inches d.b.h. and one for trees $\geq$ 11.0 inches d.b.h. The procedure outlined in Draper and Smith (1981) for fitting two linear equations with known point of intersection was used. The equation for trees < 11.0 inches d.b.h. was:

$$\text{Log } Y = a + b \text{ Log } (D^2H) + \epsilon \quad (5)$$

The equation for trees $\geq$ 11.0 inches d.b.h. was:

$$\text{Log } Y = a + b \text{ Log } (11^2H) + c \text{ Log } (D^2/11^2) + \epsilon \quad (6)$$

Researchers have used each of the models described at different times and no one model is best for all species and tree components. However, some researchers favor one model over another (McClure 1983, Clark et al., 1985).

The minimum top diameter at which trees are cut during harvest varies significantly and is dictated by local practice and logging conditions. To meet the needs of all users it is necessary for a researcher to develop a series of independently fitted equations to estimate stem weight to several top diameters (2-, 4-, 7-, 9-inch top). This approach is cumbersome and often leads to what is called crossover (Williams 1982). Crossover occurs when the line for predicted weight to a 7-inch top crosses the line for predicted weight to a 4-inch top for the same d.b.h. and height tree. This result leads to the illogical conclusion that stems cut at a 7-inch top weigh more than those logged to a 4-inch top.

To avoid this problem, researchers (Burkhart 1977, Van Deusen et al. 1981) have suggested predicting the weight of the total stem and then predicting the ratio of the stem weight to a specified top:total stem weight. We have found the exponential form of Burkharts' model, shown below, to be the best model for natural pine and hardwoods.

$$R = e^{a(d)^b (D)^c} \quad (7)$$

where: R = ratio of stem weight to specified top to weight of total stem
d = specified top diameter
D = tree d.b.h.
e = base of natural log.

This ratio model can be used to estimate stem ratios to any specified top diameter when cruising by d.b.h., d.b.h. and total height, or height to 4-inch top. These three tree dimensions are good predictors of total stem weight. Sawtimber trees, particularly hardwoods, are cruised by d.b.h. and saw-log height to estimate saw-log stem weight as accurately as possible with biomass equations. The total stem ratio model (7) is not applicable to trees cruised by d.b.h. and saw-log height. The following nonlinear model (Clark et al. 1985b) is suggested for predicting the ratio of weight to a specified top diameter:saw-log stem weight:

$$R_s = e^a \left[(Mh)^b \left(\left(1 - \left(\frac{d}{.78D} \right)^2 \right)^2 \right)^c \right] \quad (8)$$

where: R_s = weight of stem to top diameter/weight of saw-log stem ratio
Mh = saw-log merchantable height
d = specified top diameter
D = d.b.h. and 0.78D estimates the d.o.b. at the top of the first 16-foot saw log

Model (8) can be used to estimate the weight of the stem in saw-log trees to any specified diameter without crossover, assuming d is less than 0.78D.

Summary

A review of the literature shows that weight equations for entire trees and tree components are available for 11 southern softwood species and 52 southern hardwood species occurring in natural stands. The species for which equations are available make up over 92 percent of the commercial volume in the South. No specieswide or southwide biomass equations have been developed, but species equations are available for the Upland-South, the Southern Appalachian Mountains, the Piedmont, the Gulf and Atlantic Coastal Plains, and the Southeastern United States. Because of possible geographic differences in weight, only regional or locally developed and tested species equations should be used when estimating tree biomass for marketing purposes.

Literature Cited

- Art, H.W.; Marks, P.L. A summary table of biomass and net annual primary production in forest ecosystems of the world. Young, H.E., ed. In: forest biomass studies: 15th IUFRO Congress; 1971 March 15-20; Gainesville, FL. Orono, ME: University of Maine Press; 1971: 3-32.
- Baldwin, V.C. A summary of published southern pine plantation tree biomass studies. In: Baldwin, V.C., Jr.; Lohrey, R.E., eds. Proceedings of 1982 southern forest biomass working group workshop; 1982 June 16-18; Alexandria, LA; 1982: 33-45. [Available from Louisiana State University, Baton Rouge, LA]
- Belanger, Roger P. Volume and weight totals for plantation-grown sycamore. Res. Pap. SE-107. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1973. 8 pp.
- Brenneman, B.B.; Frederick, D.J.; Gardner, W.E. [and others]. Biomass of species and stands of West Virginia hardwoods. In: Proceedings of the Central Hardwood Forest Conference II; 1978 November 16; West Lafayette, IN: Purdue University; 1978: 159-170.
- Brenneman, Bruce B.; Daniels, Richard F. A case study, with alternatives, for calculating biomass yield. In: Baldwin, V.C., Jr.; Lohrey, R.E., eds. Proceedings of 1982 southern forest biomass working group workshop. 1982 June 16-18; Alexandria, LA. Baton Rouge, LA: Louisiana State University, 1982: 105-110.
- Burkhart, H.E. Cubic-foot volume of loblolly pine to any merchantable top limit. Southern Journal of Applied Forestry 1(2): 7-9; 1977.
- Burkhart, H.E.; Parker, R.C.; Oderwald, R.C. Yields for natural stands of loblolly pine. Res. Pub. FWS-2-72. Blacksburg, VA: Virginia Polytechnic Institute and State University; 1972. 63 pp.
- Carter, M.D.; White, E.H. Dry weight and nutrient accumulation in young stands of cottonwood (*Populus deltoides*, Barts.). Cir. 1970. Auburn, AL: Auburn University, Agricultural Experiment Station; 1971. 14 pp.

- Clark, Alexander; Taras, Michael A.; Schroeder, James G. Predicted green lumber and residue yields from the merchantable stem of yellow-poplar. Res. Pap. SE-119. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1974. 15 pp.
- Clark, A. III; Taras, M.A. Slash pine sawtimber stem and sawmill residue yields. Res. Pap. SE-122. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1975. 14 pp.
- Clark, A. III; Taras, M.A. Biomass of shortleaf pine in a natural sawtimber stand in northern Mississippi. Res. Pap. SE-146. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1976. 32 pp.
- Clark, Alexander; Schroeder, James G. Biomass of yellow-poplar in natural stands in western North Carolina. Res. Pap. SE-165. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1977. 41 pp.
- Clark, Alexander; Phillips, Douglas R.; Hitchcock, Harry C. Predicted weights and volumes of scarlet oak trees on the Tennessee Cumberland Plateau. Res. Pap. SE-214. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1980a. 23 pp.
- Clark, Alexander; Phillips, Douglas R.; Schroeder, James G. Predicted weights and volumes of northern red oak trees in western North Carolina. Res. Pap. SE-209. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1980b. 22 pp.
- Clark, Alexander III; Phillips, Douglas R.; Hitchcock, Harry C. III. Predicted weights and volumes of southern red oak trees on the Highland Rim in Tennessee. Res. Pap. SE-208. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1980c. 23 pp.
- Clark, Alexander; McNab, W. Henry. Total tree weight tables for mockernut, hickory, and white ash in North Georgia. Ga. For. Res. Pap. 33. Macon, GA: Georgia Forestry Commission; 1982. 11 pp.
- Clark, Alexander III. Predicting biomass production in the South. In: Hotvedt, James D.; Jackson, Ben D., eds. Proceedings of the 31st annual forestry symposium: predicting growth and yield in the mid-south; 1982 April 13-14; Baton Rouge, LA: Louisiana State University; 1982a: 119-139.
- Clark, Alexander III. Status of biomass estimation of southern hardwood species and stands. In: Baldwin, V.C., Jr.; Lohrey, R.E., eds. Proceedings of 1982 southern forest biomass working group workshop; 1982 June 16-18; Alexandria, LA; 1982b: 3-15. [Available from Louisiana State University, Baton Rouge, LA]
- Clark, Alexander III; Muse, H. David; Phillips, Douglas R.; Frederick, Douglas J. Use of segmented log-log equations to estimate tree biomass. In: Saucier, J.R., ed. Proceedings of the 1984 southern forest biomass workshop. Sixth annual meeting of the southern forest biomass working group; 1984 June 5-7; Athens, GA. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1985a: 51-57.
- Clark, Alexander III; Phillips, Douglas R.; Frederick, Douglas J. Weight, volume, and physical properties of major hardwood species in the Gulf and Atlantic Coastal Plains. Res. Pap. SE-250. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1985b. 66 pp.
- Clark, Alexander III; Saucier, J.R.; McNab, Henry. Total-tree weight, stem weight and volume tables for hardwood species in the Southeast. Ga. For. Res. Pap. 60. Macon, GA: Georgia Forestry Commission; 1986a. 44 pp.
- Clark, Alexander III; Schroeder, James G. Weight, volume, and physical properties of major hardwood species in the Southern Appalachian Mountains. Res. Pap. SE-253. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1986. 63 pp.
- Clark, Alexander III; Phillips, Douglas R.; Frederick, Douglas J. Weight, volume, and physical properties of major hardwood species in the Piedmont. Res. Pap. SE-255. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1986b. 78 pp.
- Clark, Alexander III; Phillips, Douglas R.; Frederick, Douglas J. Weight, volume, and physical properties of major hardwood species in the Upland-South. Res. Pap. SE-____. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1986c. ____ pp. [in press].
- Clark, A. III; McNab, W.H. Aboveground weight and volume of eastern white pine and hemlock in the southern Appalachians. Res. Pap. SE-____. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1986. ____ pp. [in process].
- Clark, Alexander III. Weight and volume equations for southeastern tree species [Unpublished office report]. Athens, GA: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1986. 28 pp.

- Clark, Alexander III; Phillips, Douglas R.; Frederick, Douglas J. Geographic variation of stem specific gravity, moisture content, and weight per cubic foot of southern hardwoods. *Forest Products Journal*; 1986d. [in press].
- Draper, N.R.; Smith, H. Applied regression analyses. 2nd ed. New York: John Wiley & Sons, Inc.; 1981. 709 pp.
- Edwards, M. Boyd, Jr. Weight prediction for 10 understory species in central Georgia. Res. Note SE-235. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1976. 3 pp.
- Edwards, M. Boyd, Jr.; McNab, W. Henry. Biomass prediction for young southern pines. *Journal of Forestry* 77(5): 291-292; 1979.
- Ford, Earl Carlton. Predictors of tree weights and center of gravity for various species of Appalachian hardwoods. [Unpublished]. Blacksburg, VA: Virginia Polytechnic Institute and State University; 1976. 342 pp. Thesis.
- Glover, G. Biomass tables for southern forest tree species. [Unpublished]. Coop. Agreement 19-316. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1980.
- Hitchcock, H.C. III. Above ground tree weight equations for hardwood seedlings and samplings. *TAPPI* 61(10): 119-120; 1978.
- Hitchcock, H.C.; McDonnell, J. Biomass measurement: a synthesis of the literature. In: *Forest Resource Inventories: proceedings of the workshop*. 1979 July 23-26; Fort Collins, CO: Colorado State University. Vol. II: 1979: 544-595.
- Jacobs, M.W.; Montieth, D.B. Feasibility of developing regional weight tables. *Journal of Forestry* 79(10): 676-677; 1981.
- Johansen, R.W.; McNab, W.H. Estimating logging residue weights from standing slash pine for prescribed burns. *Southern Journal of Applied Forestry* 1(2): 2-6; 1977.
- Keays, J.L. Complete tree utilization: an analysis of the literature. Inf. Rep. VP-C-69, 70, 71, 77, 79. Canada, Vancouver, B.C.: Western Forest Products Laboratory; 1971.
- King, W.W.; Schnell, Robert L. Biomass estimates of black oak tree components. Tech. Note B1. Norris, TN: Tennessee Valley Authority; 1972. 24 pp.
- Lenhart, J.D. Estimating the above-stump biomass of loblolly pine, shortleaf pine, southern red oak, and sweetgum trees in mixed pine hardwood stands in east Texas. Rep. [Unpublished]. Nacogdoches, TX: Stephen F. Austin State University, School of Forestry; 1981.
- Loomis, R.M.; Phares, R.E.; Crosby, J.S. Estimating foliage and branchwood quantities in shortleaf pine. *Forest Science* 12(1): 30-39; 1966.
- Madgwick, H.A.I. Mensuration of forest biomass. Young, H.E., ed. In: *Oslo biomass studies: 16th IUFRO Congress*; 1976 June 22; Oslo, Norway. Orono, ME: University of Maine Press; 1976: 13-27.
- Madgwick, H.A.I.; Olah, F.D.; Burkhart, H.E. Biomass of open-grown Virginia pine. *Forest Science* 23(1): 89-91; 1977.
- Madgwick, H.A.I.; Kreh, R.E. Biomass estimation for Virginia pine trees and stands. *Forest Science* 26(1): 107-111; 1980.
- Matney, T.G. Biomass equations for selected species in Oklahoma. Tech. Rep. [Unpublished]. Hot Springs, AR: Weyerhaeuser Company, Hot Springs Forestry Research Center; 1977.
- McClure, Joe P.; Schreuder, Hans T.; Wilson, Rodney L. A comparison of several volume table equations of loblolly pine and white oak. Res. Pap. SE-240. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1983. 8 pp.
- McNab, W.H.; Clark A. III. Total tree and major component green weight of white pine and hemlock in north Georgia. Ga. For. Res. Pap. 31. Macon, GA: Georgia Forestry Commission; 1982. 15 pp.
- McNab, W. Henry; Clark, Alexander III; Thomas, Charles E. Status of biomass of southern conifers in natural stands. In: Baldwin, V.C., Jr.; Lohrey, R.E., eds. *Proceedings of 1982 southern forest biomass working group workshop*; 1982 June 16-18; Alexandria, LA; 1982: 25-32. [Available from Louisiana State University, Baton Rouge, LA]
- McNab, W. Henry; Saucier, Joseph R.; Phillips, Douglas R. Green weight and wood and bark properties of pond cypress in central Florida. In: Daniels, R.F., Dunham, P.H., eds. *Proceedings of 1983 southern forest biomass working group workshop*; 1983 June 15-17; Charleston, SC; 1984: 67-71. [Available from Southeastern Forest Experiment Station, Asheville, NC]
- McNab, W.H. Total tree and product weight of beetle-killed loblolly pines. Ga. For. Res. Pap. 42. Macon, GA: Georgia Forestry Commission; 1983. 7 pp.
- McNab, W. Henry. Total tree and firewood weights and volumes of scrub oaks in the Georgia sandhills. Ga. For. Res. Pap. 22. Macon, GA: Georgia Forestry Commission; 1981. 15 pp.
- Mitsch, W.J.; Ewel, K.C. Comparative biomass and growth of cypress in Florida wetlands. *American Midland Naturalist* 101(2): 417-426;

- Nelson, L.E.; Switzer, G.L. Estimating weights of loblolly pine trees and their components in natural stands and plantations in central Mississippi. Tech. Bull. 73. Starkesville, MS: Mississippi Agricultural Forestry Experiment Station, Mississippi State University; 1975. 15 pp.
- Phillips, Douglas R.; Schroeder, James G.; Taras, Michael A. Predicted green lumber and residue yields from the merchantable stem of black oak trees. Res. Pap. SE-120. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1974. 10 pp.
- Phillips, D.R.; Schroeder, J.G. Predicted green lumber and residue yields from the merchantable stem of shortleaf pine. Res. Pap. SE-128. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Experiment Station; 1975. 12 pp.
- Phillips, Douglas R. Total-tree weights and volumes of understory hardwoods TAPPI 60(6): 68-71; 1977.
- Phillips, Douglas R. Predicted total-tree biomass of understory hardwoods. Res. Pap. SE-223. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1981. 22 pp.
- Phillips, Douglas R.; McNab, W. Henry. Total tree green weights of sapling-size pines in Georgia. Ga. For. Res. Pap. 39. Macon, GA: Georgia Forestry Commission; 1982. 19 pp.
- Phillips, Douglas R.; McNab, W. Henry. Predicted total tree biomass of sapling-size pines. Res. Pap. SE-____. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1986. ____ pp. [in press].
- Phillips, Douglas R. Status of biomass estimation of southern wood sapling-size softwoods and hardwoods. In: Baldwin, V.C., Jr.; Lohrey, R.E. eds. Proceedings of 1982 southern forest biomass working group workshop. 1982 June 16-18; Alexandria, LA; 1982: 17-23. [Available from Louisiana State University, Baton Rouge, LA]
- Reams, Gregory A.; Sullivan, Alfred D.; Matney, Thomas G.; Stevens, Robert R. Estimating above-ground biomass of slash pine and sweetgum. Tech. Bull. 110. Mississippi State, MS: Mississippi State University; 1982; 11 pp.
- Rowell, Charles E.; Carpenter, Stanley B. Black locust biomass prediction on eastern Kentucky strip mine. Southern Journal of Applied Forestry 7(1): 27-30; 1983.
- Santee, W.R. A dimensional analysis of eastern hemlock (*T. canadensis*). Athens, GA: University of Georgia; 1978. Thesis.
- Saucier, J.R.; Phillips, D.R.; Williams, J.G., Jr. Green weight, volume, board-foot, and cord tables for the major southern pine species. Ga. For. Res. Pap. 19. Macon, GA: Georgia Forestry Commission; 1981. 63 pp.
- Saucier, J.R.; Boyd, J.A. Aboveground biomass of Virginia pine in north Georgia. Res. Pap. SE-232. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1982. 32 pp.
- Saucier, Joseph R.; Clark, Alexander III. Tables for estimating total tree and product weight and volume of major southern tree species and species groups. APA-85-A-11. Washington, DC: Southwide Energy Committee, American Pulpwood Association Inc.; 1985. 59 pp.
- Saucier, Joseph R.; McNab, W. Henry. Predicted weights and volumes of live oak in northwest Florida. Res. Pap. SE-____. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1986a. ____ pp. [in process].
- Saucier, Joseph R. Predicted weights and volumes of eucalyptus in south Florida. Res. Pap. SE-____. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1986b. ____ pp. [in process].
- Schlaegel, Bryce E. Willow oak volume and weight tables for the Mississippi Delta. Res. Pap. SO-173. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1981. 14 pp.
- Schlaegel, Bryce E. Boxelder (*Acer negundo* L.) biomass component regression analysis for the Mississippi Delta. Forest Science 28(2): 355-358; 1982.
- Schlaegel, Bryce E.; Willson, Regan B. Nuttall oak volume and weight tables. Res. Pap. SO-186. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1983. 14 pp.
- Schlaegel, Bryce E. Overcup oak volume and weight tables. Res. Pap. SE-207. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1984a. 14 pp.
- Schlaegel, Bryce E. Sugarberry volume and weight tables. Res. Pap. SO-205. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1984b. 13 pp.
- Schlaegel, Bryce E. Sweetgum volume and weight tables. Res. Pap. SO-204. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1984c. 14 pp.
- Schlaegel, Bryce E. Green ash volume and weight tables. Res. Pap. SE-206. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1984d. 14 pp.

- Schnell, Robert L. Biomass estimates of hickory tree components. Tech. Note B30. Norris, TN: Tennessee Valley Authority; 1978. 39 pp.
- Schnell, R.R. Biomass estimates of eastern redcedar tree components. Tech. Note B15. Norris, TN: Tennessee Valley Authority; 1976. 15 pp.
- Schroeder, J.G.; Taras, M.A.; Clark A. III. Stem and primary product weights for longleaf pine sawtimber trees. Res. Pap. SE-139. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1975. 15 pp.
- Sirois, Donald L. Biomass of four hardwoods from lower Piedmont pine-hardwood mixed stands in Alabama. Res. Pap. SO-46. New Orleans, LA: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station; 1983. 18 pp.
- Stanek, W.; State D. Equations predicting primary productivity of trees, shrubs, and lesser vegetation based on current literature. Rep. BX-X-183. Canada, Victoria, B.C.: Canadian Forest Service, Pacific Research Center; 1978. 38 pp.
- Storey, T.G.; Fons, W.I.; Sauer, F.M. Crown characteristics of several coniferous tree species. Interim Tech. Rep. AFSW-416. Berkely, CA: U.S. Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station; 1955. 95 pp.
- Swindel, B.F.; Hollis, C.A. III; Conde, L.F.; Smith, J.E. Aboveground live biomass of slash pine trees in natural stands. IMPAC Rep. 4(1). Gainesville, FL: University of Florida, School of Forest Resources and Conservation; 1979. 16 pp.
- Swindel, B.F.; Conde, Louis F.; Smith, Joel E.; Hollis, Charles A. III. Green weights of major tree species in north Florida pine flatwoods. Southern Journal of Applied Forestry 3(2): 74-78; 1982.
- Taras, M.A.; Schroeder, J.G.; Phillips, D.R. Predicted green lumber and residue yields from the merchantable stem of loblolly pine. Res. Pap. SE-121. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1974. 11 pp.
- Taras, M.A.; Clark, A. III. Aboveground biomass of loblolly pine in a natural, uneven-aged sawtimber stand in central Alabama. TAPPI 58(2): 103-105; 1975.
- Taras, M.A.; Clark, A. III. Aboveground biomass of longleaf pine in a natural sawtimber stand in southern Alabama. Res. Pap. SE-162. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1976. 32 pp.
- Taras, M.A.; Phillips, D.R. Aboveground biomass of slash pine in a natural sawtimber stand in southern Alabama. Res. Pap. SE-188. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1978. 31 pp.
- Taras, M.A. Aboveground biomass of Choctawhatchee sand pine in northwest Florida. Res. Pap. SE-210. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1980. 23 pp.
- Tritton, Louise M.; Hornbeck, James W. Biomass equations for major tree species of the northeast. Gen. Tech. Rep. NE-69. Broomall, PA: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1982. ___ pp. [in press].
- U.S. Department of Agriculture, Forest Service. Tables of whole-tree weight for selected U.S. tree species. Gen. Tech. Rep. WO-42. Washington, DC; 1984. 157 pp.
- Van Deusen, P.C.; Sullivan, A.D.; Matney, T.G. Prediction system for cubic-foot volume of loblolly pine applicable through much of its range. Southern Journal of Applied Forestry 5(4): 186-189; 1981.
- Wade, D.D. Estimating slash quantity from standing loblolly pine. Res. Note SE-125. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1969. 4 pp.
- Walters, Earl O. Tree taper and green weight equations for east Texas southern yellow pine in natural stands. In: Baldwin, V.C., Jr.; Lohrey, R.E. eds. Proceedings of 1982 southern forest biomass working group workshop; 1982 June 16-18; Alexandria, LA; 1982b: 71-80. [Available from Louisiana State University, Baton Rouge, LA]
- Wendel, G.W. Fuel weights of pond pine crowns. Res. Note SE-149. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1960. 2 pp.
- Whittaker, R.H.; Cohen, H.; Olson, J.S. Net production relations of three tree species at Oak Ridge, Tennessee. Ecology 44(4): 806-810; 1963.
- Wiant, H.V.; Scheetz, C.E.; Colominno, A. [and others]. Tables and procedures for estimating weights of some Appalachian hardwoods. Bull. 659T. Morgantown, WV: West Virginia University, Agricultural and Forestry Experiment Station; 1977. 36 pp.
- Wiant, Harry V.; Castanedo, Froylan; Sheetz, Carter E. [and others]. Equations for predicting weights of some Appalachian hardwoods. WV For. Note No. 7. Morgantown, WV: West Virginia University; 1979. 21-26.

Williams, James G. Modeling problems in predicting total-tree component biomass. In: Baldwin, V.C., Jr.; Lohrey, R.E. eds. Proceedings of 1982 southern forest biomass working group workshop; 1982 June 16-18; Alexandria, LA; 1982: 111-115. [Available from Louisiana State University, Baton Rouge, LA]

Young, H.E. A summary and analysis of weight table studies. Young, H.E., ed. In: Oslo Biomass Studies: 16th IUFRO Congress: 1976 June 22; Oslo Norway. Orono, ME: University of Maine Press; 1976: 251-282.

Table 1.--Summary of total tree and tree component weight equations for southern hardwoods ≥ 1.0 inches d.b.h. by species and location

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
Number Inches					
Ash, green (<i>Fraxinus pennsylvanica</i> Marsh.)					
S. and S.E. Coastal Plain	158	1-18	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1985b
Mississippi Delta	70	1-30	DTh	TT,ST	Schlaegel, 1984d
Ash, white (<i>Fraxinus americana</i> L.)					
Northeast Georgia	31	5-18	DH4,DMh	TT,SST,ST4 (G)	Clark & McNab, 1982
North Georgia and Western North Carolina	52	5-22	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & Schroeder, 1986
S.E. West Virginia	15	2-16	D	TT (G&D)	Brenneman et al., 1978
S.E. West Virginia	13	2-16	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Basswood, American (<i>Tilia americana</i> L.)					
Western North Carolina	18	6-16	D,DTh,DH4	TT,ST/R (G&D)	Clark & Schroeder, 1986
S.E. West Virginia	13	6-19	D	TT (G&D)	Benneman et al., 1978
S.E. West Virginia	19	5-19	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Beech, American (<i>Fagus grandifolia</i> Ehrh.)					
S.E. West Virginia	56	2-15	D	TT (G&D)	Brenneman et al., 1978
S.E. West Virginia	24	5-15	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Birch, sweet (<i>Betula lenta</i> L.)					
S.E. West Virginia	8	4-15	D	TT (G&D)	Brenneman et al., 1978
Western North Carolina	21	6-18	D,DTh,DH4	TT,ST/R (G&D)	Clark & Schroeder, 1986
S.E. West Virginia	21	6-16	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Birch, yellow (<i>Betula alleghaniensis</i> Britton)					
S.E. West Virginia	24	2-16	D	TT (G&D)	Brenneman et al., 1978
Blackgum (<i>Nyssa sylvatica</i> var. <i>biflora</i> (Walt.) Sarg.)					
S. and S.E. Coastal Plain	135	1-20	D,DTh,DH4,DMh	TT,ST/R,SST (G&D)	Clark et al., 1985b
S.E. United States	135	1-20	DTh,DH4,DMh	TT,ST/R,SST (G)	Clark et al., 1986a
Central Georgia	20	1- 6	BD	TT (D)	Edwards, 1976
Blackgum (<i>Nyssa sylvatica</i> Marsh.)					
Western North Carolina	18	6-18	D,DTh,DH4	TT,ST/R (G&D)	Clark & Schroeder, 1986
Western North Carolina	18	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Tennessee	15	1- 2	DTh,Th,DBTh	TT (G&D)	Hitchcock, 1978
Boxelder (<i>Acer negundo</i> L.)					
Mississippi Delta ^{3/}	49	1-14	DTh	TT,ST4,CR (G&D)	Schlaegel, 1982

Continued

Table 1.--Summary of total tree and tree component weight equations for southern hardwoods ≥ 1.0 inches d.b.h. by species and location--Continued

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
Number Inches					
Buckthorn, Carolina (<i>Rhamnus caroliniana</i> Walt.)					
Tennessee	15	1- 2	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Cherry, black (<i>Prunus serotina</i> Ehrh.)					
Northern West Virginia	21	2-16	D	TT,ST4,CR (G&D)	Wiant et al., 1977
Northern West Virginia	21	2-16	DTh	TT,ST4,CR (G&D)	Wiant et al., 1979
S.E. West Virginia	24	5-16	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
S.E. West Virginia	26	3-16	D	TT (G&D)	Brenneman et al., 1978
Tennessee	17	1- 2	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Cherry, pin (<i>Prunus pensylvanica</i> L. f.)					
S.E. West Virginia	20	7-12	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Cottonwood, eastern (<i>Populus deltoides</i> Bartr. ex Marsh.)					
Alabama	48	1-11	D	TT (D)	Carter & White, 1971
Dogwood, (<i>Cornus florida</i> L.)					
Western North Carolina	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Piedmont Georgia	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Western North Carolina	24	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Piedmont Southeast	24	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Tennessee	16	1- 2	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Central Georgia	20	1- 6	BD	TT (D)	Edwards, 1976
Elm, American (<i>Ulmus americana</i> L.)					
Piedmont of Southeast	16	1-11	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1986b
Eucalyptus, grandis (<i>Eucalyptus grandis</i>)					
South Florida	28	2-12	DTh	TT,ST/R,CR (G&D)	Saucier, 1986b
Eucalyptus, robusta (<i>Eucalyptus robusta</i>)					
South Florida	28	2-12	DTh	TT,ST/R,CR (G&D)	Saucier, 1986b
Hickory, mockernut (<i>Carya tomentosa</i> (Poir.) Nutt.)					
Central Alabama	18	4- 9	DTh	TT,ST4,CR (G&D)	Sirois, 1983
North Georgia	27	5-19	DH4,DMh	TT,SST,ST4 (G)	Clark & McNab, 1982
Hickory, shagbark (<i>Carya ovata</i> (Mill.) K. Koch)					
S.E. West Virginia	14	3-17	D	TT (G&D)	Brenneman et al., 1978
S.E. West Virginia	13	7-17	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Hickory, species (<i>Carya</i> sp.)					
S. and S.E. Coastal Plain	42	1-18	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1985b
Piedmont of Southeast	22	1-18	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1986b
S. Appalachian Mountains	54	5-23	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & Schroeder, 1986
Mid-South	37	1-18	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1986c

Continued

Table 1.--Summary of total tree and tree component weight equations for southern hardwoods ≥ 1.0 inches d.b.h. by species and location--Continued

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
	<u>Number</u>	<u>Inches</u>			
Hickory, species (<i>Carya sp.</i>) (con't.)					
S.E. United States	64	1-19	DTh,DH4,DMh	TT,SST,ST/R (G)	Clark et al., 1986a
Eastern Oklahoma	22	1-14	DTh	TT,ST4 (G&D)	Matney, 1977
Eastern Tennessee	37	2-28	D	TT,SST,CR (G&D)	Schnell, 1978
Northern West Virginia	19	2-16	DTh	TT,ST4,CR (G&D)	Wiant et al., 1979
Northern West Virginia	19	2-16	D	TT,ST,CR (G&D)	Wiant et al., 1977
Western North Carolina	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Piedmont Georgia	11	1- 5	DTh	TT (G&D)	Phillips, 1977
Western North Carolina	24	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Piedmont Southeast	22	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Tennessee	16	1- 3	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Loblolly-bay (<i>Gordonia lasianthus</i> (L.) Ellis)					
North Florida	19	2- 9	D	ST (G)	Swindel et al., 1982
Locust, black (<i>Robinia pseudoacacia</i> L.)					
S. Appalachian Mountains	18	6-16	D,DTh,DH4	TT,ST/R (G&D)	Clark & Schroeder, 1986
Eastern Kentucky	1,371	1- 5	DTh	TT (D)	Rowell & Carpenter, 1983
Magnolia, cucumber (<i>Magnolia acuminata</i> L.)					
S.E. West Virginia	6	8-16	D	TT (G&D)	Brenneman et al., 1978
Maple, red (<i>Acer rubrum</i> L.)					
S. and S.E. Coastal Plain	85	1-15	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1985b
Piedmont of Southeast	32	1-16	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1986b
S. Appalachian Mountains	30	5-17	D,DTh,DH4	TT,ST/R (G&D)	Clark & Schroeder, 1986
Western Virginia	95	1-16	DTh,DH4	TT,ST4 (G)	Ford, 1976
Northern West Virginia	20	2-16	DTh	TT,ST4,CR (G&D)	Wiant et al., 1979
Northern West Virginia	20	2-16	D	TT,ST4,CR (G&D)	Wiant et al., 1977
S.E. West Virginia	27	2-17	D	TT (G&D)	Brenneman et al., 1978
S.E. West Virginia	13	5-17	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Western North Carolina	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Piedmont Georgia	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Western North Carolina	24	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Piedmont Southeast	23	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Tennessee	16	1- 3	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Maple, sugar (<i>Acer saccharum</i> Marsh.)					
S.E. West Virginia	119	2-15	D	TT (G&D)	Brenneman et al., 1978
S.E. West Virginia	38	5-15	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Oak, black (<i>Quercus velutina</i> Lam.)					
Western North Carolina	40	12-26	DMh	SST (G)	Phillips et al., 1974
S. Appalachian Mountains	26	6-22	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & Shroeder, 1986
Tennessee	26	12-35	D	TT,SST,CR (G&D)	King & Schnell, 1972
Northern West Virginia	22	2-16	DTh	TT,ST4,CR (G&D)	Wiant et al., 1979
Northern West Virginia	22	2-16	D	TT,ST4,CR (G&D)	Wiant et al., 1977
Oak, blackjack (<i>Quercus marilandica</i> Muenchh.)					
Oklahoma	15	1-12	DTh	TT,ST4,CR (G&D)	Matney, 1977

Continued

Table 1.--Summary of total tree and tree component weight equations for southern hardwoods ≥ 1.0 inches d.b.h. by species and location--Continued

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
	<u>Number</u>	<u>Inches</u>			
Oak, chestnut (<i>Quercus prinus</i> L.)					
S. Appalachian Mountains	50	6-22	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & Schroeder, 1986
Piedmont of Southeast	37	1-15	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1986b
Western Virginia	125	1-17	DTh,DH4	TT,ST4,CR (G)	Ford, 1976
Northern West Virginia	21	2-16	D	TT,ST4,CR (G&D)	Wiant et al., 1977
Northern West Virginia	21	2-16	DTh	TT,ST4,CR (G&D)	Wiant et al., 1979
S.E. West Virginia	13	3-13	D	TT (G&D)	Brenneman et al., 1978
S.E. West Virginia	12	5-13	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Western North Carolina	11	1- 5	DTh	TT (G&D)	Phillips, 1977
Western North Carolina	24	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Tennessee	15	1- 2	DTh,Th,BDTh	TT (G&D)	Hitchcock, 1978
Oak, laurel (<i>Quercus laurifolia</i> Michx.)					
S. and S.E. Coastal Plain	47	1-17	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1985b
Oak, laurel and water combined (<i>Q. laurifolia</i> Michx. and <i>Q. nigra</i> L.)					
S.E. United States	148	1-20	DTh,DH4,DMh	TT,SST,ST/R (G)	Clark et al., 1986a
Oak, live (<i>Quercus virginiana</i> Mill.)					
Northwest Florida	37	3-30	D,DTh,DH4,DMh	TT,SST,ST/R (G)	Saucier & McNab, 1986
Oak, nuttall (<i>Quercus nuttallii</i> Palmer)					
Mississippi Delta	56	3-38	DTh	TT,ST4 (G&D)	Schlaegel & Willson, 1983
Oak, overcup (<i>Quercus lyrata</i> Walt.)					
Mississippi Delta	88	1-34	DTh	TT,ST4,CR (G&D)	Schlaegel, 1984a
Oak, post (<i>Quercus stellata</i> Wangenh.)					
Piedmont Southeast	12	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Mid-South	26	5-21	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1986c
Oak, northern red (<i>Quercus rubra</i> L.)					
S. Appalachian Mountains	71	6-24	D,DTh,DH4,DMh	TT,SST,ST4,CR(G&D)	Clark et al., 1980b
S. Appalachian Mountains	71	6-24	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & Schroeder, 1986
Northern West Virginia	22	2-16	DTh	TT,ST4,CR (G&D)	Wiant et al., 1979
Northern West Virginia	22	2-16	D	TT,ST4,CR (G&D)	Wiant et al., 1977
S.E. West Virginia	24	4-21	D	TT (G&D)	Brenneman et al., 1978
S.E. West Virginia	23	5-21	DTh	TT,ST4 (G)	Brenneman & Daniels, 1982
Oak, scarlet (<i>Quercus coccinea</i> Muenchh.)					
Piedmont Southeast	32	5-18	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1986b
S. Appalachian Mountains	27	6-22	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & Schroeder, 1986
Mid-South	42	5-20	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1986c
Central Tennessee	28	5-20	D,DTh,DH4,DMh	TT,SST,ST4,CR(G&D)	Clark et al., 1980a
Western Virginia	43	2-20	DTh,DH4	TT,ST4 (G)	Ford, 1976
Northern West Virginia	20	2-16	DTh	TT,ST4,CR (G&D)	Wiant et al., 1979
Northern West Virginia	20	2-16	D	TT,ST4,CR (G&D)	Wiant et al., 1977

Continued

Table 1.--Summary of total tree and tree component weight equations for southern hardwoods ≥ 1.0 inches d.b.h. by species and location--Continued

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
	<u>Number</u>	<u>Inches</u>			
Oak, scrub (<i>Quercus</i> sp.)					
Sand Hills Georgia	124	1- 7	D, DTh	TT, ST (G)	McNab, 1981
Oak, southern red (<i>Quercus falcata</i> Michx.)					
Central Alabama	25	3-12	DTh	TT, ST, CR (G&D)	Sirois, 1983
Piedmont Southeast	48	2-19	D, DTh, DH4, DMh	TT, SST, ST/R (G&D)	Clark et al., 1986b
Oklahoma	16	1-18	DTh	TT, ST4, CR (G&D)	Matney, 1977
Central Tennessee	29	5-22	D, DTh, DH4, DMh	TT, SST, ST4, CR (G&D)	Clark et al., 1980c
East Texas	33	2-21	DTh	TT, ST4, CR (G&D)	Lenhart, 1981
Mid-South	60	5-22	D, DTh, DH4, DMh	TT, SST, ST/R (G&D)	Clark et al., 1986c
Piedmont Georgia	13	1- 5	DTh	TT (G&D)	Phillips, 1977
Piedmont Southeast	25	1- 5	D, DTh	TT (G&D)	Phillips, 1981
Oak, southern red and scarlet combined (<i>Q. alba</i> L. and <i>Q. coccinea</i> Muenchh.)					
S.E. United States	98	5-22	DTh, DH4, DMh	TT, SST, ST/R (G)	Clark et al., 1986a
Oak, water (<i>Quercus nigra</i> L.)					
S. and S.E. Coastal Plain	112	1-20	D, DTh, DH4, DMh	TT, SST, ST/R (G&D)	Clark et al., 1985b
North Florida	10	3-13	D	ST (G)	Swindel et al., 1982
Oak, white (<i>Quercus alba</i> L.)					
Central Alabama	14	4- 9	DTh	TT, ST, CR (G&D)	Sirois, 1983
S.E. United States	128	1-22	DTh, DH4, DMh	TT, SST, ST/R (G)	Clark et al., 1986a
S. and S.E. Coastal Plain	38	6-20	D, DTh, DH4, DMh	TT, SST, ST/R (G&D)	Clark et al., 1985b
Piedmont Southeast	44	2-21	D, DTh, DH4, DMh	TT, SST, ST/R (G&D)	Clark et al., 1986b
S. Appalachian Mountains	28	5-22	D, DTh, DH4, DMh	TT, SST, ST/R (G&D)	Clark & Schroeder, 1986
Mid-South	51	1-20	D, DTh, DH4, DMh	TT, SST, ST/R (G&D)	Clark et al., 1986c
Oklahoma	32	1-18	DTh	TT, ST4, CR (G&D)	Matney, 1977
Western Virginia	70	1-20	DTh, DH4	TT, ST4 (G)	Ford, 1976
Northern West Virginia	21	2-16	DTh	TT, ST4, CR (G&D)	Wiant et al., 1979
Northern West Virginia	21	2-16	D	TT, ST4, CR (G&D)	Wiant et al., 1977
S.E. West Virginia	29	3-14	D	TT (G&D)	Brenneman et al., 1978
S.E. West Virginia	28	5-14	DTh	TT, ST4 (G)	Brenneman & Daniels, 1982
Western North Carolina	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Piedmont Georgia	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Western North Carolina	18	1- 5	D, DTh	TT (G&D)	Phillips, 1981
Piedmont Southeast	23	1- 5	D, DTh	TT (G&D)	Phillips, 1981
Tennessee	16	1- 3	DTh, Th, BDTh	TT (G&D)	Hitchcock, 1978
Oak, willow (<i>Quercus phellos</i> L.)					
Mississippi Delta	79	2-37	DTh	TT, ST4, (G&D)	Schlaegel, 1981
Sugarberry (<i>Celtis laevigata</i> Willd.)					
Mississippi Delta	--	--	DTh	TT, ST4, CR (G&D)	Schlaegel, 1984b
Sweetbay (<i>Magnolia virginiana</i> L.)					
North Florida	6	2- 9	D	ST (G)	Swindel et al., 1982
Persimmon (<i>Diospyros virginiana</i> L.)					
Tennessee	19	1- 3	DTh, BDTh, Th	TT (G&D)	Hitchcock, 1978

Continued

Table 1.--Summary of total tree and tree component weight equations for southern hardwoods ≥ 1.0 inches d.b.h. by species and location--Continued

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
<u>Number</u> <u>Inches</u>					
Sassafras (<i>Sassafras albidum</i> (Nutt.) Nees.)					
Tennessee	17	1- 3	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Serviceberry (<i>Amelanchier</i> sp.)					
Tennessee	12	1- 2	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Sourwood (<i>Oxydendrum arboreum</i> (L.) DC)					
Western North Carolina	11	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Tennessee	17	1- 3	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Sumac sp. (<i>Rhus</i> sp.)					
Tennessee	16	1- 2	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978
Sweetgum (<i>Liquidambar styraciflua</i> L.)					
Central Alabama	15	4-10	DTh	TT,ST,CR (G&D)	Sirois, 1983
Coastal Plain	313	1-20	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1985b
Piedmont Southeast	236	1-21	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1986b
Mid-South	39	2-17	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1986c
Mississippi Delta	54	1-32	DTh	TT,ST/R (G&D)	Schlaegel, 1984c
Mississippi	60	1-20	DTh	TT,SST,ST4,CR(G&D)	Reans et al., 1982
East Texas	30	4-12	DTh	TT,SST,ST4,CR(G&D)	Lenhart, 1981
S.E. United States	263	1-21	DTh,DH4,DMh	TT,SST,ST/R (G)	Clark et al., 1986a
S.E. United States	263	1-21	DTh,DH4,DMh	TT,SST,ST/R (G)	Saucier & Clark, 1985
Piedmont Georgia	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Piedmont Southeast	24	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Central Georgia	20	1- 6	BD	TT (D)	Edwards, 1976
Sycamore (<i>Platanus occidentalis</i> L.)					
Central Georgia	103	4-10	DTh	ST (G&D)	Belanger, 1973
Piedmont Southeast	29	2-20	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1986b
Tupelo, water (<i>Nyssa aquatica</i> L.)					
Coastal Plain Alabama	--	1-20	D,DTh,DH4,DMh	TT,SST,ST4,CR(G&D)	Glover, 1980
S. and S.E. Coastal Plain	79	1-20	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1985b
Yellow-poplar (<i>Liriodendron tulipifera</i> L.)					
S. and S.E. Coastal Plain	26	5-21	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1985b
S. Appalachian Mountains	62	6-26	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & Schroeder, 1986
Piedmont Southeast	78	3-20	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark et al., 1986b
Mid-South	19	5-19	D,DTh,DH4	TT,ST/R (G&D)	Clark et al., 1986c
Western North Carolina	39	6-28	DTh	TT,SST,ST4,CR(G&D)	Clark & Schroeder, 1977
Western North Carolina	47	12-28	DMh	SST,ST4 (G&D)	Clark et al., 1974
Northern West Virginia	21	2-16	DTh	TT,ST4,CR (G&D)	Wiant et al., 1979
Northern West Virginia	21	2-16	D	TT,ST4,CR (G&D)	Wiant et al., 1977
S.E. West Virginia	12	6-15	D	TT (G&D)	Brenneman et al., 1978
S.E. United States	117	1-22	DTh,DH4,DMh	TT,SST,ST/R (G)	Saucier & Clark, 1985
S.E. United States	117	1-22	DTh,DH4,DMh	TT,SST,ST/R (G)	Clark et al., 1986a

Continued

Table 1.--Summary of total tree and tree component weight equations for southern hardwoods ≥ 1.0 inches d.b.h. by species and location--Continued

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
	<u>Number</u>	<u>Inches</u>			
Yellow-poplar (<i>Liriodendron tulipifera</i> L.) (con't.)					
Western North Carolina	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Piedmont Georgia	12	1- 5	DTh	TT (G&D)	Phillips, 1977
Western North Carolina	24	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Piedmont Southeast	24	1- 5	D,DTh	TT (G&D)	Phillips, 1981
Tennessee	17	1- 2	DTh,BDTh,Th	TT (G&D)	Hitchcock, 1978

^{1/}D = tree stem diameter at breast height; Th = tree height; H4 = height to 4-inch d.o.b. top; Mh = height to saw-log merchantable height; BD = basal diameter.

^{2/}TT = total tree above stump; SST = saw-log stem; ST4 = stem from butt to 4-inch d.o.b. top; ST = stem from butt to top; ST/R = stem from butt to top with ratio equations; CR = crown, G = green weight; D = dry weight.

^{3/}Metric units.

Table 2.--Summary of total tree and tree component weight equations for southern softwoods ≥ 1.0 inches d.b.h. by species and location

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
	<u>Number</u>	<u>Inches</u>			
Pine, loblolly (<i>Pinus taeda</i> L.)					
Georgia	448	6-18	DTh	TT,ST/R (G)	Saucier et al., 1981
Central Alabama	41	6-20	DTh	TT,SST,ST4,CR (G&D)	Taras & Clark, 1975
Central Georgia	18	6-17	D	CR (D)	Wade, 1969
Central Georgia	60	5-18	DMh,DH4	TT,SST,ST4 (G&D)	McNab, 1983
Central Mississippi	36	2-20	DTh	TT,ST4,CR (G&D)	Nelson & Switzer, 1975
Eastern Oklahoma	81	1-24	DTh	TT,ST4,CR (G&D)	Matney, 1977
Southern South Carolina	48	10-20	DMh	SST (G)	Taras et al., 1974
Southern South Carolina	9	1-20	D	CR (D)	Storey et al., 1955
East Texas	38	2-24	DTh	TT,ST4,CR (G&D)	Lenhart, 1981
Eastern Virginia	721	1-18	DTh	ST4 (G&D)	Burkhart et al., 1972
S.E. United States	913	1-24	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark, 1986
Piedmont Georgia	25	1- 4	BDTh	TT (D)	Edwards & McNab, 1979
Georgia	100	1- 5	D,DTh	TT (G&D)	Phillips & McNab, 1986
East Texas	359	5-23	DTh,DH4	ST/R (G)	Walters, 1982

Pine, loblolly and shortleaf (*Pinus taeda* L. and *P. echinata* Mill.)

Piedmont Southeast	1026	1-20	DTh,DH4,DMh	TT,SST,ST/R (G)	Saucier & Clark, 1985

Pine, loblolly, shortleaf, and virginia (*Pinus taeda* L., *P. echinata* Mill., *P. virginia* Mill.)

Piedmont Georgia	200	1- 5	D,DTh	TT (G)	Phillips & McNab, 1982

Continued

Table 2.--Summary of total tree and tree component weight equations for southern softwoods ≥ 1.0 inches d.b.h. by species and location--Continued

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
	<u>Number</u>	<u>Inches</u>			
Pine, shortleaf (<i>Pinus echinata</i> Mill.)					
Georgia	228	6-18	DTh	TT,ST/R (G)	Saucier et al., 1981
Northern Mississippi	34	6-20	DTh	TT,SST,ST4,CR (G&D)	Clark & Taras, 1976
Northern Mississippi	57	10-20	DMh	SST,ST4, (G)	Phillips & Schroeder, 1975
Southeast Missouri	182	6-16	D	CR (D)	Loomis et al., 1966
Eastern Oklahoma	81	1-24	DTh	TT,ST4,CR (G&D)	Matney, 1977
Eastern Tennessee	10	2-11	D	CR (D)	Whittaker, 1963
East Texas	38	2-24	DTh	TT,TST,CR (G&D)	Lenhart et al., 1981
S.E. United States	467	1-20	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark, 1986
Piedmont Georgia	22	1- 4	BDTh	TT (D)	Edwards & McNab, 1979
Georgia	100	1- 5	D,DTh	TT (G&D)	Phillips & McNab, 1986
East Texas	242	1-20	DTh,DH4	ST/R	Walters, 1982
Pine, slash (<i>Pinus elliotii</i> Engelm. var. <i>Elliotii</i>)					
Georgia	208	6-18	DTh	TT,SST,ST/R (G)	Saucier et al., 1981
Southern Alabama	43	6-20	DTh	TT,SST,ST4,CR (G&D)	Taras & Phillips, 1978
Southern Alabama	43	9-20	DMh	SST,ST4 (G)	Clark & Taras, 1975
North Florida	128	2-14	D	TT,TST,CR (G)	Swindel et al., 1979
North Florida	132	1-14	D	TST (G)	Swindel et al., 1982
South Georgia	16	5-11	D	CR (D)	Johansen & McNab, 1977
S.E. United States	435	1-21	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark, 1986
Coastal Plain Georgia	29	1- 4	BDTh	TT (D)	Edwards & McNab, 1979
Georgia	80	1- 5	D,DTh	TT (G&D)	Phillips & McNab, 1986
Pine, longleaf (<i>Pinus palustris</i> Mill.)					
South Georgia	273	6-18	DTh	TT,ST/R (G)	Saucier et al., 1981
South Alabama	48	10-20	DMh	SST (G)	Schroeder et al., 1975
South Alabama	47	6-18	DTh	TT,SST,ST4,CR (G)	Taras & Clark, 1976
S.E. United States	552	1-19	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark, 1986
Coastal Plain Georgia	24	1- 4	BDTh	TT (D)	Edwards & McNab, 1979
Georgia	80	1- 5	D,DTh	TT (G&D)	Phillips & McNab, 1986
East Texas	169	1-20	DTh,DH4	ST/R	Walters, 1982
Pine, loblolly, slash & longleaf (<i>Pinus taeda</i> L., <i>P. elliotii</i> Engelm. var. <i>elliotii</i> , <i>P. palustris</i> Mill.)					
Coastal Plain Southeast	1285	1-24	DTh,DH4,DMh	TT,SST,ST/R (G&D)	Saucier & Clark, 1985
Coastal Plain Georgia	200	1- 5	D,DTh	TT (G)	Phillips & McNab, 1982
Pine, virginia (<i>Pinus virginiana</i> Mill.)					
North Georgia	25	6-14	DTh	TT,ST/R (G&D)	Saucier et al., 1981
Central Virginia	324	--	DTh	TST,CR (D)	Madgwick & Kreh, 1980
Southwest Virginia	14	--	Th	TST,CR (D)	Madgwick et al., 1977
North Georgia	25	6-14	DTh,DH4	TT,ST/R (G&D)	Clark, 1986
Georgia	80	1- 5	D,DTh	TT (G&D)	Phillips & McNab, 1986
Pine, virginia, white, & shortleaf (<i>Pinus virginiana</i> Mill., <i>P. strobus</i> L., <i>P. echinata</i> Mill.)					
S. Appalachian Mountains	100	1- 5	D,DTh	TT (G)	Phillips & McNab, 1982
Pine, sand (<i>Pinus clausa</i> var. <i>immuginata</i> Ward.)					
Northwest Florida	36	4-15	DTh	TT,SST,ST4,CR (G&D)	Taras, 1980

Continued

Table 2.--Summary of total tree and tree component weight equations for southern softwoods ≥ 1.0 inches by species and location--Continued

Species and location	Sample size	D.b.h. range	Independent variables ^{1/}	Dependent variables ^{2/}	Investigator
	<u>Number</u>	<u>Inches</u>			
Pine, pond (<i>Pinus serotina</i> Michx.)					
Eastern North Carolina	20	3-15	DTh	CR (D)	Wendel, 1960
Pine, eastern white (<i>Pinus strobus</i> L.)					
North Georgia	36	6-18	DH4,DMh	TT,SST,ST/R (G)	McNab & Clark, 1982
North Georgia	36	6-18	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & McNab, 1986
North Georgia	40	1- 5	D,DTh	TT (G&D)	Phillips & McNab, 1986
Hemlock, eastern (<i>Tsuga canadensis</i> (L.) Carr.)					
North Georgia	36	6-18	DH4,DMh	TT,SST,ST/R (G)	McNab & Clark, 1982
North Georgia	36	6-18	D,DTh,DH4,DMh	TT,SST,ST/R (G&D)	Clark & McNab, 1986
Western North Carolina	12	1-10	D	TT,TST,CR (D)	Santee, 1978
S.E. West Virginia	21	--	D	TT (G&D)	Brenneman et al., 1978
Red cedar, eastern (<i>Juniperus virginiana</i> L.)					
Eastern Tennessee	31	6-17	D	TT,TST,CR (G&D)	Schnell, 1976
Pondcypress (<i>Taxodium distichum</i> var. <i>nutans</i> (Ait.) Sweet)					
Central Florida	65	5-18	DTh	TT,TST (G&D)	McNab et al., 1984
North Florida	51	1-18	D	TST (G)	Swindel et al., 1982
Central Florida	10	3-16	DTh	TT,TST,CR (D)	Mitsch & Ewel, 1979

^{1/}D = tree stem diameter at breast height; Th = tree height; H4 = height to 4-inch d.o.b. top; Mh = height to saw-log merchantable height; BD = basal diameter.

^{2/}TT = total tree above stump; SST = saw-log stem; ST4 = stem from butt to 4-inch d.o.b. top; ST = stem from butt to top; ST/R = stem from butt to top with ratio equations; CR = crown, G = green weight; D = dry weight.

METHODS FOR ESTIMATING THE FOREST BIOMASS IN TENNESSEE VALLEY REGION

J. Daniel Thomas and Robert T. Brooks, Jr.

Forest Resource Analyst and Project Leader,
Respectively
Forest Resource Inventory and Analysis Project
Tennessee Valley Authority
Norris, TN 37828-2000

Species-specific weight prediction equations utilizing a segmented log-log form were applied to U.S. Forest Service plot- and tree-level data from State inventories in the 201-county Tennessee Valley region to produce the detailed biomass estimates essential for regional resource management and industrial development. The forest biomass inventory; annual growth, removals, and net change of merchantable growing stock biomass; and the biomass potentially available each year for use as energywood were estimated. The total oven-dry weight of aboveground woody biomass is estimated to be 1,223.3 million tons or 40.2 tons per acre of commercial forest land.

Introduction

Following a general realization in the 1970s of the implications of their dependence on other countries as energy suppliers, Americans began a search for alternative energy sources. One of the first alternatives to be utilized was America's original, abundant, renewable resource--biomass fuels--primarily wood. The forests of the Tennessee Valley region, having recovered from near devastation in the early part of the century were believed to be ready to meet part of future energy demands.

When foresters were asked to quantify the available resource for energy, it was soon realized that insufficient information existed for estimating the weight of trees; and a major effort was directed toward developing accurate methods of estimating forest biomass. The results of this effort are only now becoming fully available.

The methods presented here combined with the most recent and complete data were used to produce the detailed forest biomass estimates essential to comprehensive, regional resource management and TVA's industrial development program. These estimates are available in customized reports for any group of counties in the Tennessee Valley region and a surrounding buffer zone through the Forest Resources Information System (FRIS).

Geographic Area

The 201-county Tennessee Valley region consists of the 125 counties in the Tennessee River Drainage Basin plus 76 counties outside the Basin served by TVA power. These counties include the entire State of Tennessee plus portions of the

States of Alabama, Georgia, Kentucky, Mississippi, North Carolina, and Virginia.

Methods

Data Sources

Plot- and tree-level forest inventory data were obtained through cooperative agreements with the U.S. Forest Service Experiment Stations serving the Tennessee Valley region. The Southeastern Forest Experiment Station provided the data for Georgia (1982), North Carolina (1984), and Virginia (1977). The Southern Forest Experiment Station provided data for Alabama (1982), Tennessee (1980), and Mississippi (1977). Data for Kentucky (1975) were provided by the Northeastern Forest Experiment Station. These were the most recent forest inventories available. The average inventory date for the Tennessee Valley region is 1980.

County-level data were summarized from the plots within each county inventoried by the Southeastern and Southern Forest Experiment Stations. For the Kentucky counties, data from the new-ground plots were used to derive county-level estimates based on survey-unit partitioning techniques.

Biomass Estimation

Species-specific weight prediction equations were used to estimate green and oven-dry weights of wood and wood and bark in the bole, crown, total stem, and total tree (Clark et al., in press-a; Clark et al., in press-b; Clark, personal communication, 1985). These equations utilize a segmented log-log equation form. Two equations were fit simultaneously using a known point of intersection (11.0 inches dbh for hardwoods and 9.0 inches dbh for softwoods). Statistical comparison indicated that a segmented log-log method of deriving equations was much more accurate than use of a single log-log equation. The logarithmic forms of the equations were converted to original units by applying a correction factor to the antilogarithm to remove the bias incurred by estimating the geometric mean rather than the arithmetic mean. An exponential ratio equation was used to estimate the proportion of predicted total stem weight to a specified top diameter outside bark.

Tree weight equations were available for species comprising 83 percent of the cubic foot volume of hardwood species and 92 percent of the softwood species in the region. For the remaining species, equations were not available. Substitutions of existing equations were made based on specific gravity and crown form.

The segmented equations which utilize merchantable height and diameter at breast height (dbh) were applied to each tree. Weight estimates were derived for growing stock, rough and rotten trees, small (understory) trees, and growth and removals of growing stock. Net change was computed by subtracting removal weight from growth weight.

Total tree weight

In this paper, biomass is defined as the weight of all wood and bark (excluding foliage) above a 1-foot stump in all live trees that are 1.0 inch or greater in diameter at breast height (4.5 feet aboveground) and located on commercial forest land.

The following equations were used to estimate total tree weight for growing stock, rough and rotten trees, and understory trees:

For trees less than 5.0 inches dbh and greater than 1.0 inch dbh:

$$(1) \text{ TTW} = a * (\text{DBH}^2 * \text{Ht})^b$$

where:

TTW = Total tree weight

a and b = Species-specific coefficients for green or oven-dry wood or wood and bark

Ht = Total height

DBH = Diameter at breast height

For hardwoods less than 11.0 inches dbh and greater than or equal to 5.0 inches dbh and softwoods less than 9.0 inches dbh and greater than or equal to 5.0 inches dbh:

$$(2) \text{ TTW} = c * (\text{DBH}^2 * \text{H4})^d$$

where:

TTW = Total tree weight

c and d = Species-specific coefficients for green or oven-dry wood or wood and bark

H4 = Height to a 4.0 inch top outside bark

DBH = Diameter at breast height

For hardwoods greater than or equal to 11.0 inches dbh and softwoods greater than or equal to 9.0 inches dbh:

$$(3) \text{ TTW} = f * (\text{DBH}^2)^g * \text{H4}^i$$

where:

TTW = Total tree weight

f, g, and i = Species-specific coefficients for green or oven-dry wood or wood and bark

H4 = Height to a 4-inch top outside bark

DBH = Diameter at breast height

Bole and sawlog weight

To obtain the bole weight of an individual tree, equations (2) and (3) were applied using appropriate coefficients to solve for total stem weight. The results were multiplied by a bole-stem ratio to give the weight of the merchantable bole to a 4-inch top. If the tree qualified as a sawlog, the total stem weight was multiplied by a sawlog-stem ratio to yield the sawlog weight.

The following equation form was used to calculate bole-stem and sawlog-stem ratios:

$$(4) R_x = e^{(j * \text{DOB}^k) * \text{DBH}^m}$$

where:

R_x = Bole-stem ratio or sawlog-stem ratio

e^x = Base of natural logarithm (2.71828)

j, k, and m = Species specific coefficients for green or oven-dry wood or wood and bark

DOB = Diameter outside bark

DOB = 4.0 inches for all boles

7.0 inches for softwood sawlogs

9.0 inches for hardwood sawlogs

DBH = Diameter at breast height

Thus bole weight was obtained by:

$$(5) \text{ BW} = \text{TSW} * R_B$$

where:

BW = Bole weight

TSW = Total stem weight

R_B = Bole-stem ratio

and sawlog weight (where trees qualified) was calculated by:

$$(6) \text{ SLW} = \text{TSW} * R_S$$

where:

SLW = Sawlog weight

TSW = Total stem weight

R_S = Sawlog-stem ratio

Crown and pulpwood weight

Once the values of total tree and bole were known, crown and pulpwood weights were calculated by subtraction as follows.

$$(7) \text{ CW} = \text{TTW} - \text{BW}$$

and

$$(8) \text{ PW} = \text{BW} - \text{SLW}$$

where:

CW = Crown weight

TTW = Total tree weight

BW = Bole weight

PW = Pulpwood weight

SLW = Sawlog weight

Growth weight

Growth weights were obtained by first computing the weight per cubic foot for the bole of each tree and then multiplying this weight by the cubic foot growth volume to give the growth weight in the bole.

$$(9) \text{ GWB} = (\text{BW}/\text{CBV}) * \text{GVOL}$$

where:

GWB = Growth weight of the bole

BW = Bole weight

CBV = Current bole volume

GVOL = Growth volume

Total tree growth weight was then calculated based on a ratio of bole growth weight to total bole weight.

$$(10) \text{ TTGW} = (\text{GWB}/\text{BW}) * \text{TTW}$$

where:

TTGW = Total tree growth weight

GWB = Growth weight of the bole

BW = Bole weight

TTW = Total tree weight

Total stem growth weight was calculated in the same manner by substituting total stem weight for total tree weight in equation (10). Sawlog growth weight was computed by replacing total stem weight with total stem growth weight in equation (6). Likewise the respective growth weights were substituted in equations (7) and (8) to obtain crown and pulpwood growth weights.

Removal weight calculation

Removal weights were calculated using the past dbh and past height to a 4-inch top in equation (2) or (3) to yield past weight of the total tree and past weight of the total stem. Current total stem weight was then calculated using a past weight to past volume ratio times the current volume.

$$(11) \text{ TSW} = (\text{PTSW}/\text{PVOL}) * \text{CVOL}$$

where:

TSW = Total stem weight
PTSW = Past total stem weight
PVOL = Past volume
CVOL = Current volume

The current bole-stem ratio was determined by using current dbh in equation (4). This ratio was used with the result of (11) in equation (5) to produce current bole weight of removal trees. Total tree weight was computed by replacing past total stem weight in equation (11) with past total tree weight. Crown and pulpwood weights were derived as they were for growing stock trees with equations (7) and (8). Sawlog weights were computed using the appropriate sawlog ratio from (4).

Potential annual energywood

The weight of wood and bark available for energywood can be described as the sum of three components.

1. The weight of net change available (including crowns) minus the weight of sawlogs (reserved for conventional forest products).
2. The portion of rough and rotten inventory including crowns, which could be removed in conjunction with the conventional harvest of net change.
3. Logging residues from current harvests (crowns).

The estimate of potential annual energywood can be considered as the sum of future energywood harvests plus logging residues from current harvests. The components of future energywood harvests are the total weight of pulpwood size trees (5.0 inches dbh up to sawlog size), the portion of sawtimber tree boles above the sawlog plus the crowns of sawlog trees, and the total weight of rough and rotten or cull trees harvested. The sawlog portion of sawtimber trees is excluded since it likely could be marketed for a higher return. Only the portion of the rough and rotten

inventory which would be removed in connection with the harvest of the available net change is included in the energywood estimates. This quantity is estimated by multiplying the growing stock net change rate of the major species group by the rough and rotten inventory.

Assuming a market for energy wood exists, loggers currently harvesting timber for other products could remove the tops of trees which they are currently leaving in the woods. These are logging residues from current harvests. While logging residues are defined as including remaining rough and rotten trees, crowns, and other residual material, only the crowns are considered here resulting in a more conservative estimate.

The weight of potential annual energywood can be expressed mathematically as follows:

$$(12) \text{ PAEW} = ((G - R) - (SG - SR)) + (RR * ((G - R)/GSI)) + LR \\ = (GSNC - SNC) + (RR * GSNC) + LR \\ = \text{NCA} + \text{RRA} + \text{LR}$$

where:

PAEW = Potential Annual Energywood
G = Annual Growth of Growing Stock
R = Annual Removals of Growing Stock
SG = Annual Sawlog Growth
SR = Annual Sawlog Removals
RR = Rough and Rotten Inventory
GSI = Growing Stock Inventory
LR = Logging Residues
GSNC = Growing Stock Net Change
SNC = Sawlog Net Change
NCA = Net Change Available
RRA = Rough and Rotten Available

Standard Errors

Standard errors were computed by source (growing stock, rough and rotten, understory) and major species group (hardwood, softwood, total) for each county and for all counties in a State using the standard error function in a statistical computing package where N was the number of plots in a county or State. These were then expressed as a percent of the mean.

Due to the sampling techniques used in Kentucky, standard errors were computed using a slightly different method. The total number of new ground plots was determined for each county. The variance for each variable for a survey unit was calculated and divided by the number of plots in each county in the unit to give the county variance. The square root of the county variance was expressed as a percent of the county. The standard error for all Kentucky counties in the region was similarly computed.

These methods gave standard errors comparable to those in traditional forest inventory. Contribution to the error by the equations used to compute biomass was ignored. Table 1 shows the standard errors in percent for the Tennessee Valley region.

Table 1. Standard Errors for Oven-dry Aboveground Woody Biomass by Source and Major Species Group for the Tennessee Valley Region.

Species group	Growing stock	Rough and rotten	Understory	Total
-----Percent-----				
Softwood	2.0	3.5	3.5	2.0
Hardwood	3.1	2.5	3.4	3.0
Total	2.9	2.5	3.3	2.8

Results

The total oven-dry weight of aboveground woody biomass in the Tennessee Valley region is estimated to be 1,223.3 million tons (± 2.8 percent) covering 30.4 million acres of commercial forest land or 51 percent of the total land area. This is equivalent to 40.2 oven-dry tons per acre of commercial forest land.

The following are key statistics for the region:

- Growing stock trees contribute 894.0 million oven-dry tons. Rough and rotten or cull trees add 161.6 million oven-dry tons, and understory trees add 167.7 million oven-dry tons to the total.
- The majority of the standing inventory (1005.6 million oven-dry tons or 82 percent) consists of hardwoods. Softwoods account for 217.7 million oven-dry tons or 18 percent of the inventory.

Table 2. Oven-dry Weight of Aboveground Woody Biomass by Source and Major Species Group for the Tennessee Valley Region.

Species group	Growing stock	Rough and rotten	Understory	Total
-----Million tons-----				
Softwood	185.8	7.2	24.7	217.7
Hardwood	708.2	154.4	143.0	1005.6
Total	894.0	161.6	167.7	1223.3

- The greatest softwood per acre weights occur in the Mississippi, Alabama, and Georgia portion of the region, while hardwood per acre highs are in North Carolina and Virginia.
- Private landowners hold 75 percent of the total biomass inventory. Forest industry owns 8 percent and national forests and other public holdings contain 17 percent.

- Annual growth exceeds removals in merchantable growing stock--increasing inventory at a rate of 2.2 percent per year.
- There are 18.8 million tons (0.62 ton per acre) of oven-dry energywood potentially available in the Tennessee Valley region each year. This is the energy equivalent of 0.27 quads or more than one third of the energy derived in 1985 from coal at TVA's steam plants (Mills In press).
- Softwood energy wood totals 2.8 million tons (0.09 ton per acre) while hardwoods contribute 16.0 million tons (0.53 ton per acre).
- The greatest amounts of potential annual energywood are located in the Virginia counties (0.84 ton per acre) while Alabama counties have the least (0.41 ton per acre).

Summary

Species-specific weight prediction equations utilizing a segmented log-log form can be applied traditional forest inventory data to produce regional biomass estimates across U.S. Forest Service Experiment Station boundaries. Customized reports for any group of counties in the Tennessee Valley region and a surrounding buffer zone are available through the Forest Resources Information System.

Literature Cited

- Clark, A. III; D. R. Phillips; D. J. Frederick. Weight, volume, and physical properties of major hardwood species in the Gulf and Atlantic Coastal Plains. Res. Pap. SE-250. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; 1985.
- Clark, A. III; J. G. Shroeder. Weight, volume, and physical properties of major hardwood species in the Southern Appalachian Mountains. Res. Pap. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station; In Press-a.
- Clark, A. III; D. R. Phillips; D. J. Frederick. Weight, volume, and physical properties of major hardwood species in the Piedmont. Res. Pap. Asheville, NC: U.S. Department of Agriculture, Southeastern Forest Experiment Station; In Press-b.
- Mills, Debra D., ed. TVA handbook-1986. Knoxville, TN: Tennessee Valley Authority, Technical Library. In Press.

AREAS OF BIOMASS RESEARCH

Boris Zeide

Professor of Forestry, Department of Forest Resources University of Arkansas at Monticello, Agricultural Experiment Station, Monticello, AR 71655, USA.

Biomass research is classified into areas according to the major forms of biomass heterogeneity produced by the spatial and temporal discreteness of trees. It is suggested that, at present, ecologically-oriented areas deserve equal or higher priority than the technical aspects of biomass studies.

Introduction

Although biomass research has been the major accomplishment in forest science during the last two decades, its potential contribution far exceeds the present achievements which are largely confined to the estimation of existing biomass. In order to answer some practically urgent problems, future biomass studies should address, along with development of more efficient methods for accurate biomass estimations, the problems of predicting the natural course of biomass dynamics and the effects of management activities.

The wide and expanding scope of biomass studies makes it necessary to look at the organization of our efforts and to classify them into more or less uniform research areas. Such a classification establishes a framework for the discussion of previous works and the guidance of future studies. To be meaningful and convenient, a classification should be based on some of the important attributes of forest biomass.

One of these pivotal attributes is biomass variability. Biomass estimation would have been quite simple if biomass had covered the Earth with a complete and homogeneous layer of a fixed depth. Actual biomass distribution differs from this simplest uniformity in many forms. Of them, three major forms, described below, should be considered as separate areas of biomass research.

Much of biomass variability can be traced to the fact that living matter exists in the form of discrete individuals and not as an undivided whole. Discovery of other, hidden units of matter discreteness (cell, gene, atom, etc.) required a great deal of effort and profoundly affected the corresponding branches of science and our knowledge as a whole. Feynman et al. (1963, p. 1-2) wrote at the beginning of "The Feynman lectures on physics": "If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generations of creatures, what statement would contain the most information in the fewest words? I believe it is the atomic hypothesis (or the

atomic fact, or whatever you wish to call it) that all things are made of atoms - little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. In that one sentence, you will see, there is an enormous amount of information about the world, if just a little imagination and thinking are applied."

In contrast, the existence of discrete organisms is too apparent to be acknowledged and is usually overlooked in the lists of ecological concepts, principles and laws (McIntosh 1980). Yet the discreteness on the organism level might be as fruitful as that on other levels of matter organization and, in particular, helpful in structuring biomass research and better understanding forest ecosystems.

Organism discreteness means that there are two basic variables for overall description of living creatures: their mass and their number per unit area. Accordingly, in addition to ecological taxonomy, there are two main branches of ecology: production ecology and population ecology (Harper 1977) which emphasize, respectively, the mass and the number. The relationship between these variables links together the branches and belongs to neither of them; it constitutes the core of ecology. This relationship determines the change in biomass of closed stands: as trees grow, biomass per unit area increases too, despite the decline of the number of trees.

Organisms are discrete not only in space but in time as well. Fallen trees create gaps which cannot be instantly covered by neighbors' growth. Affected by growth and density-dependent mortality of individual trees, as well as by various disturbances, forest canopy is always patchy and often incomplete. This density variation, a direct result of spatial and temporal discreteness in biomass, is a second source of variability in biomass distribution. In even-aged stands the proportion of canopy closure predictably decreases with age after the thicket stage. One of the mechanisms responsible for this trend is that the size of a gap created by the fall of a single tree increases as trees become older and larger, while the ability of the neighbors to close the gap decreases with age. Stand density is of great practical concern: manipulation of tree number, which is the most manageable stand variable, is a chief method in the maximization of the end product of foresters' activities, usable tree mass.

The third kind of biomass diversity has evolved in forest communities where organism discreteness has unleashed competition for light which in turn has resulted in the highly heterogeneous structure of each tree. Tree biomass consists of several components with drastically different properties. Crown occupies the entire aboveground space allotted to a tree but constitutes only about one-fifth of its mass. Roots are also widely spread and constitute an even smaller proportion of biomass

than crowns. Stem, on the other hand, is the densest tree part which accounts for about two-thirds of tree mass with practically no utilized space of own. Investigation of relations between the components, which vary with site, species, age, density and many other factors, might greatly improve biomass estimations.

The lack of the simplest uniformity does not mean that biomass distribution is chaotic. There are certain patterns in the relationships between size and number of trees, tree structure, and gap dynamics. Investigation of these patterns can provide the basis for estimation, prediction and manipulation of forest biomass.

Along with the three mentioned ecologically-oriented areas, sampling techniques, methods of measurements and computational procedures should always remain a subject of biomass research.

Thus the following four areas capture the major directions in biomass research:

- (1) tree morphology;
- (2) stand density;
- (3) size and number of trees;
- (4) field and computational methods.

Tree Morphology

While present biomass models have been developed for the practical purpose of predicting the mass of trees or their portions, they often fail to consider the mechanical and ecological considerations affecting tree form. These factors can be helpful in biomass modeling. For instance, fundamental principles of mechanics immediately reveal that the product of squared diameter and height, a commonly used variable in biomass equations, is not a good predictor of crown mass because crown mass (as any load atop a column of equal resistance) is inversely related to tree height while it is directly proportional to the diameter. In relation to trees this fact has been often documented when crown dimensions were regressed on both diameter and height (Briegleb 1952; Madgwick and Kreh 1980; Moeur 1981; Schmitt and Grigal 1981; Alban and Laidly 1982; Ouellet 1983, 1985; Harding and Grigal 1985; Lohrey 1985). This example shows that deductive inference can be useful in biomass research which currently relies mostly on approximations without clear biological or mechanical meaning.

Ecological investigations could produce interpretations of the coefficients of biomass models in terms of ecological variables like species tolerance, crown class, stand density, and so on and, as a result, facilitate the estimation of the coefficients. Using the immense amount of biomass data collected during the last two decades, it is also possible to scrutinize some tenets of silvicultural textbooks such as the connections between tolerance, crown length and crown density (crown mass per unit of crown volume.)

Stand Density

Estimation and prediction of forest biomass can be made on the basis of relatively stable and consistent factors such as species composition and site quality. At the same time biomass is affected by numerous accidental forces such as natural disturbances which reduce stand density. The actual density of stands is always below the maximum or potential value due to the time lag between the practically instant creation of gaps by natural factors or management activities and the filling of them. Therefore, prior to investigation of the intrinsic relationship between size and number of trees, one has to take into account stand density.

Relative measures of stand density are more useful than absolute ones because the same absolute value, depending on specific conditions, might represent quite different levels of density. Relative density is the ratio of the existing biomass (or related variables) to its maximum possible on a given site. Determining this maximum value is crucial for the solution of two central forestry problems, the problem of density and stocking and the problem of site quality evaluation.

So far these problems have been either avoided by choosing variables insensitive to density (height of dominant trees) or substituted for by measuring the most dense "fully stocked" areas outside of the given stand. These indirect approaches are not always adequate. The chief example of this avoidance is site index classification. It is based on the height of dominant trees which, although less sensitive than other variables, still changes with density. Additionally, the correspondence between height and biomass is not exact. The drawbacks of the substitution approach are that there is no guarantee that the fully stocked areas are indeed maximally dense, and that the site quality of these areas and the studied stand is identical.

Therefore, what is needed now is to develop a direct approach for the estimation of the biomass limit which is required as a reference level for the calculation of relative density of stands and evaluation of site quality. The point is to determine the limit for a given stand within the stand itself, that is, to develop an intrinsic measure of density stress. This can be done by analyzing the distribution of a suitable absolute density measure estimated from plots within a given stand. This distribution must be negatively skewed because it is bounded by the density limit. When a distribution function, which reflects this asymmetry and contains a coefficient representing the limit, is fitted to the actual data, the limit can be computed even when none of the plots is fully stocked.

Size and Number of Trees

The relationship between the size of trees and their number per unit area is another area

of biomass research essential for the efficient estimation of stand biomass. Two well-known models, the Reineke (1933) equation and the $3/2$ power law of self-thinning (Yoda et al. 1963), provide a background for the further development necessitated by some limitations of the models. Among these limitations are the restricted set of variables, the trend in coefficients which are supposed to be constant (this indicates the deficiency of the models' form), and, particularly for the law, low accuracy.

According to the self-thinning law, biomass of trees is a function of one variable, the number of trees per unit area, and the relative rate of self-thinning with respect to biomass growth is a universal constant. The law could be true in two cases: (1) when all factors of stand dynamics unaccounted for by the law (in particular, change in plant form and change in crown closure) are nonexistent; (2) when the effects of the unaccounted factors on the rate of self-thinning cancel each other. Investigation of the actual effects of these neglected factors is a worthy subject of biomass research. It seems (Zeide 1985) that developmental changes in plant form and in crown closure indeed oppositely affect the rate of self-thinning. In general, however, these factors rarely balance each other and the rate predictably changes with age, species, and site quality, and does not merely fluctuate about a constant value. The limiting line of self-thinning does not have any constant slope (on the log-log scale); generally, this line is a curve. A more realistic model of self-thinning should be more inclusive and in particular it should reflect the change in crown closure, or gap dynamics. This problem deserves special attention since knowledge of this process might lead to optimization of a major forestry operation, management thinning. The analysis of the law suggests the use of tree size as a predictor of their number, the utilization of the fact that diameter is better correlated with crown width and tree number than tree mass, and the development of a more adequate form of expressing allometric relationships than the elementary power function.

Field and Computational Methods

This indispensable area is the most developed one since it is the only one traditionally considered as a proper area of biomass studies. Still, there are serious doubts about the accuracy of biomass estimates and the reliability of methods used in biomass studies. Numerous discrepancies and contradictions were reported in the literature. It is not uncommon in biomass studies to find that a tree portion is greater than the entire tree or that the whole is not equal to the sum of its parts. Different field methods frequently produce different estimates of the same stand. Reviewing the existent techniques used for biomass estimation, Parde (1980) writes that they lead to alarming inconsistencies even when the methods are close and the experimentation is careful. Another noted researcher, Cunia

(1979, p. 662) expressed a similar concern: "A review of the presently used methodology of selecting sample trees for biomass tables construction shows that seldom, if ever, the samples of trees were representative of the population of interest. Most of the time, a subjective opinion of what seems representative, random and efficient superseded statistical considerations. For this reason, serious doubts are cast over the applicability of most if not all presently constructed tree biomass tables."

Instead of routine application of known methods and regressions, perhaps, new ecologically meaningful models should be formulated which would allow the drafting of appropriate inferences and predictions amendable for testing, analysis, and exclusion of alternatives which in their turn should become a subject of continuous doubt, checking, and developing. In other words, explicit and consistent application of the scientific method is required in biomass measurements just as in any other branch of science. As soon as we stop questioning our approaches and methods, we can be pretty confident that it won't take long to go astray and produce erroneous and misleading results.

Priorities

So far the rapid progress in biomass studies has substantially increased the amount of information about forest stands but contributed little to their understanding. In biomass studies, physical efforts (cutting and weighing of trees) far exceed mental exploits which are usually limited to regressing the weights on accessible variables, chiefly, stem diameter at breast height. Except for a few ecological investigations, the studies do not set forward any hypotheses to be tested. Selection of regression equations is based on narrow statistical considerations rather than on biological reasoning. Discussions of the values of coefficients, their correspondence to theoretical expectations and relationship to species characteristics and environmental conditions are rare.

This problem is further aggravated because, with few exceptions, it is not recognized among foresters involved in biomass estimations. A notable exception to this rule is H. A. I. Madgwick, one of the most respected and experienced students of forest biomass, who wrote in his preface to the English edition of *Forest Biomass* (Satoo 1982): "Lord Rutherford has said that all science is either physics or stamp collecting. On that basis the study of forest biomass must be classified with stamp collecting and other such pleasurable pursuits".

The crucial issue of biomass studies is not a deficiency in a particular weighing technique, sampling method, or regression model. The issue is more general: it is our approach to research, the current imbalance between measurement and thinking. We keep

accumulating redundant and often incorrect data instead of learning from our own experience. Therefore, at present, the ecologically-oriented areas deserve equal or higher priority than the technical aspects of biomass studies.

To support the thesis that "with different forms of measurement as well as with improvement measurement [sic] our scientific knowledge will increase", Young and Ribe (1983, p. 158) cite the famous words by Lord Kelvin in which science is virtually equated with measuring. This was a popular idea in the latter part of the past century, but today, a remark by Lord Rutherford, Kelvin's successor at Cambridge, is more appropriate. He is said to have interrupted one of his students with the comment: "All the time I see you measuring something. It's good. But when do you take time and think?"

Literature Cited

- Alban, D. H., and P. R. Laidly. 1982. Generalized biomass equations for jack and red pine in the Lake States. *Can. J. For. Res.* 12:913-921.
- Briegleb, P. A. 1952. An approach to density measurement in Douglas-fir. *Journal of Forestry* 50:529-536.
- Cunia, T. 1979. On sampling trees for biomass tables construction: some statistical comments. In *Forest Resource Inventories*, pp. 615-628. Edited by W. E. Frayer. Department of Forest and Wood Sciences. Colorado State University, Fort Collins, Colorado. 1037 p.
- Feynman, R. P., R. B. Leighton, and M. Sands. 1963. *The Feynman lectures on physics*. Volume 1. Addison-Wesley, New York.
- Harding, R. B., and D. F. Grigal. 1985. Individual tree biomass estimation equations for plantation-grown white spruce in northern Minnesota. *Can. J. For. Res.* 15:738-739.
- Harper, J. L. 1977. The contribution of terrestrial plant studies to the development of the theory of ecology. In *Changing scenes in the life sciences, 1776-1976*, pp. 139-157. Edited by C. E. Goulden. Special Publ. 12. Acad. Nat. Sci. Phila., Penna.
- Lohrey, R. E. 1985. Aboveground biomass of planted and direct-seeded slash pine in the West Gulf Region. In *The 1984 Southern Forest Biomass Workshop*, pp. 75-82. Edited by J. R. Saucier. USDA For. Serv., Southeastern Forest Experiment Station. 121 p.
- Madgwick, H. A. I., and R. E. Kreh. 1980. Biomass estimation for Virginia pine trees and stands. *Forest Science* 26:107-111.
- McIntosh, R. P. 1980. The background and some current problems of theoretical ecology. *Synthese* 43:195-255.
- Moeur, M. 1981. Crown width and foliage weight of Northern Rocky Mountain conifers. USDA For. Serv. Res. Pap. INT-283. 14 p.
- Ouellet, D. 1983. Biomass prediction equations for twelve commercial species in Quebec. Canadian Forestry Service Information Report LAU-X-62E. 34 p.
- Ouellet, D. 1985. Biomass equations for six commercial tree species in Quebec. *Forestry Chronicle* 61:218-222.
- Parde, J. 1980. Forest biomass. *Forestry Abstracts* 41:343-362.
- Reineke, L. H. 1933. Perfecting a stand-density index for even-aged forests. *J. Agric. Res.* 46:627-638.
- Satoo, T. 1982. Forest biomass. Edited and revised by H. A. I. Madgwick. Martinus Nijhoff/Dr. W. Junk Publishers, The Hague. 152 p.
- Schmitt, M.D.C., and D.F. Grigal. 1981. Generalized biomass estimations for *Betula papyrifera* Marsh. *Can. J. For. Res.* 11:837-840.
- Yoda, K., T. Kira, H. Ogawa, and K. Hozumi. 1963. Self-thinning in overcrowded pure stands under cultivated and natural conditions. *J. Biol. Osaka City Univ.* 14:107-129.
- Young, H. E., and J. H. Ribe. Quantifying forest science with biomass. Part one: site productivity. In *Planning, performance and evaluation of growth and yield studies*, pp. 157-165. Edited by H. L. Wright. Commonwealth Forestry Institute Occasional Paper 20. Oxford. 175 p.
- Zeide, B. 1985. Tolerance and self-tolerance of trees. *Forest Ecology and Management* 13:149-166.

RESEARCH PAPERS

Biomass Studies Outside the United States

Moderator: James T. Bones

PREDICTION ERROR IN TREE BIOMASS REGRESSION

FUNCTIONS FOR WESTERN CANADA

T. Singh

Research Scientist, Northern Forestry Centre,

Canadian Forestry Service, Edmonton, Alberta

T6H 3S5

Abstract

Biomass equations were derived for eight commercial tree species in the Boreal Forest Region of western Canada, using two regression functions based on diameter, and a combination of diameter and height. The model based on diameter and height regression function gave excellent predictions when validated using independent data; R^2 obtained for the predictions based on this regression function for regional and Alberta applications were 0.99 and 0.97 for trembling aspen, and balsam fir had R^2 values of 0.96 for both the regional and Alberta applications.

Introduction

Biomass equations are increasingly being used for deriving forest biomass tables and inventories. Individual species equations are available also to provide biomass information pertaining to all or part of a tree as a renewable energy source. Extensive biomass data have been accumulated in North America over recent years for most of the forest tree species, however, little attention has been given to assessment of estimation errors and bias inherent in the use of biomass equations for such predictions.

Estimation errors can occur due to many sources. Sampling methods used for collection of data, choice of a regression function on which the prediction equation is to be based, and the lack of fit contribute to most of these errors. The precision and accuracy of prediction equations need to be known to avoid misleading interpretation.

A study was undertaken in the Boreal Forest Region of western Canada to determine and compare errors of estimation when prediction equations derived from two different regression functions are used to obtain biomass information. The objectives of the study were as follows:

1. Determine the accuracy of prediction equations as derived from individual prairie provinces (Alberta, Saskatchewan, and Manitoba) and the combined provinces (regional) data sets.
2. Compare the errors of prediction when equations from an individual province are used in other prairie provinces.
3. Assess the accuracy and bias inherent in the use of prediction equations based on two different regression functions when validated using an independent data set in Alberta.

Methods

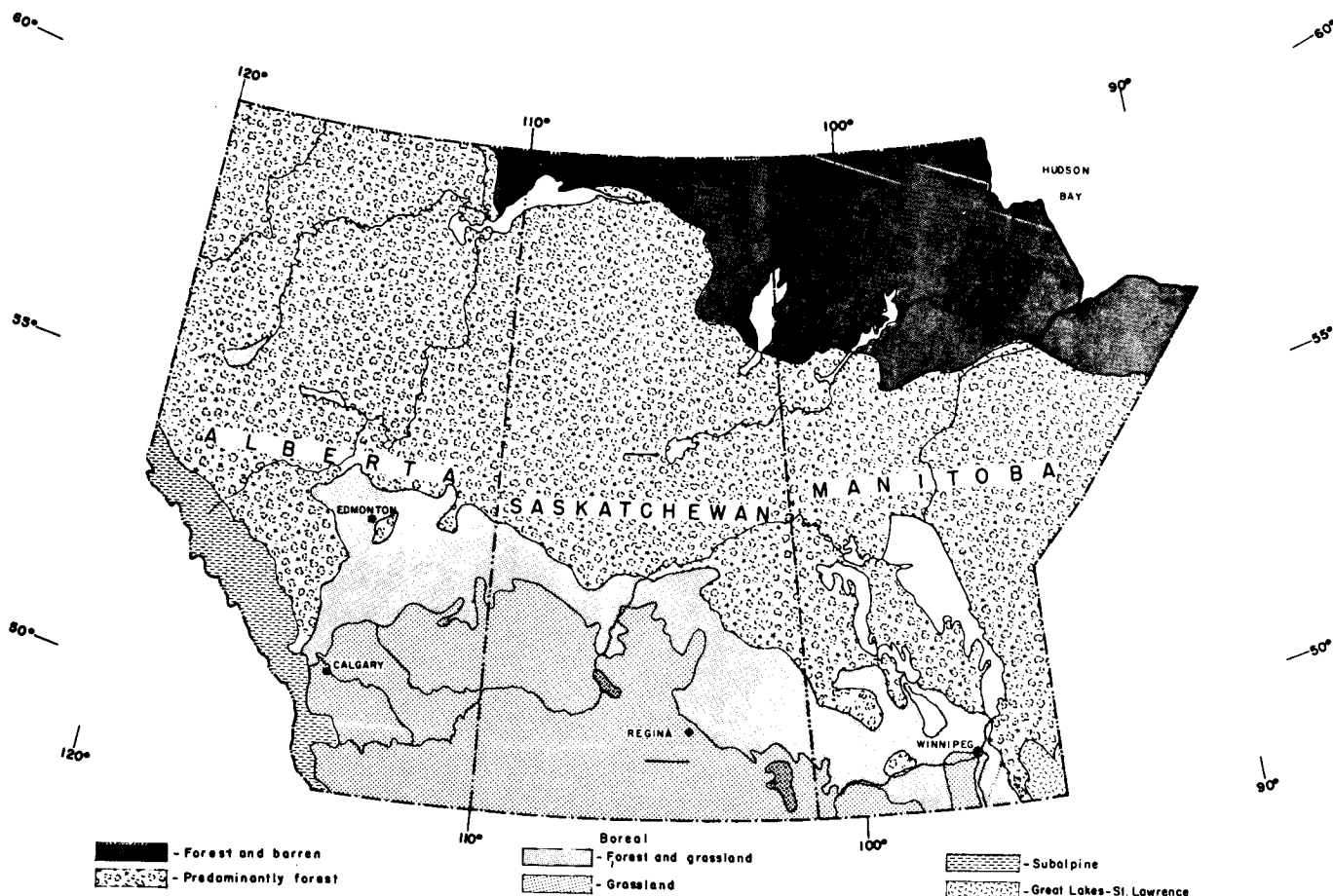
Field and Laboratory Procedures

The tree species tested in the study were five softwoods and three hardwoods. The softwoods were jack pine (*Pinus banksiana* Lamb.), black spruce (*Picea mariana* (Mill.) BSP), white spruce (*Picea glauca* (Moench) Voss), balsam fir (*Abies balsamea* (L.) Mill.), and tamarack (*Larix laricina* (du Roi) K. Koch). The three hardwoods were trembling aspen (*Populus tremuloides* Michx.), balsam poplar (*P. balsamifera* Marsh.), and white birch (*Betula papyrifera* Marsh.).

Twenty trees of each species were sampled in each of the three prairie provinces (Alberta, Saskatchewan, and Manitoba), for a total of 60 trees per species on a regional basis for the predominantly Boreal Forest Region in western Canada (Fig. 1). The field information recorded included diameter at breast height outside bark (dbhob) at 1.3 m (D), and total height of the tree (H). Sampling was done according to diameter size, by felling five trees per species per province for each of the four diameter classes (0-11, 11-20, 21-30, and 31+ cm).

After felling the trees, merchantable (dbhob 10 cm or greater) and nonmerchantable (dbhob less than 10 cm) sections of the tree stem were marked. The merchantable stem was cut into four and the nonmerchantable stem into three equal subsections. Length and dbhob measurements were made for each of the subsections, and the tree stem was cut and weighed for fresh weights of individual components. These weights were also determined for the remaining aboveground parts of the felled tree. Subsamples consisting of 1-cm thick disks were obtained for the tree stem sections, and representative subsamples were similarly taken for foliage and live branches. Debarking to separate wood and bark was done for all subsamples, except for live small branches. Oven-dry weights were determined after all subsamples were dried at 103°C for 24 hours or until a constant weight was attained. Details on field and laboratory procedures are provided by Singh (1982).

Figure 1. Boreal Forest Region in the prairie provinces of Alberta, Saskatchewan, and Manitoba.



Computer Procedures

Dry/fresh weight and bark/wood dry weight ratios were computed from the volume-weighted data. The

the Smalion formula:

$$V = L(A_t + A_b)/2$$

where L = length of section,

A_t = cross-sectional area at top, and

A_b = cross-sectional area at bottom.

The cross-sectional areas were computed from dbhob measurements. Total volume of the tree stem was obtained by summing the volumes of all component stem sections. Computer subroutines were written in FORTRAN for biomass computations based on field and laboratory data collected in the study (Singh and Campbell 1983).

Regression Functions and Analysis

Two regression functions were used, one based on a combination of D and H, and the other based on D only:

$$\text{Model I: } W = a_0 + a_1 D^2 H$$

$$\text{Model II: } W = a_0 + a_1 D + a_2 D^2 + a_3 D^3$$

where W = oven-dry biomass of live tree above-ground, and $a_0, a_1, \dots, a_3$ are regression coefficients.

Mean, standard deviation, R^2 , SEE (standard error of estimate), % SEE (standard error expressed as percentage of the mean), and residuals (difference between the actual and the fitted value) were calculated for the derived versus the applied regression equations.

Comparisons were made for testing the accuracy of the individual province and regional prediction equations using the data collected for the study and an independent data set available for two hardwood species in Alberta.

Results and Discussions

The R^2 values for the individual province equations for the eight tree species ranged from 0.883 (white birch) to 0.997 (trembling aspen) for Model I, and from 0.934 to 0.997 for Model II (Appendix I). Both species for the highest and the lowest R^2 values were sampled in Saskatchewan. The corresponding range of % SEE values was 35.2 to 5.3, and 28.0 to 6.4, respectively. The regional equations had a R^2 range of 0.943 to 0.988, with the same species in Saskatchewan making up the highest and the lowest ranking. Black spruce (Model I) also showed the highest fit in Alberta ($R^2 = 0.997$, % SEE = 5.1).

The regional equations based on the combined sample of 60 trees per species showed that the R^2 ranged from 0.943 to 0.988 for Model I, and from

0.956 to 0.977 for Model II (Appendix II). The corresponding % SEE values ranged from 24.2 to 11.2, and from 21.8 to 15.8, respectively. White birch had the lowest and trembling aspen the highest fit in this case as well. The goodness of fit and the associated errors seemed to have averaged out when the three data subsets from the prairie provinces were combined into a single regional set for deriving biomass prediction equations for each species.

The prediction and errors involved in the individual province equations when used in each province are shown in Table 1. A test of the prairies regional equations on the individual province data for each species showed that the provincial equations were slightly better than the regional equations in predicting the actual biomass values (Table 2).

Table 1. Regional equations applied to estimation of oven-dry biomass of eight tree species in three individual provinces^a

Species	Province	Actual	Predicted Mean		Predicted R ²		% SEE		Mean Residuals	
		Mean	Model I	Model II	Model I	Model II	Model I	Model II	Model I	Model II
Softwoods:										
Jack pine	Alberta	157.5	162.9	174.7	0.97	0.95	17.8	23.0	-5.4	-17.1
	Saskatchewan	189.1	192.5	178.8	0.98	0.97	12.8	18.0	-3.4	10.3
	Manitoba	197.9	189.0	190.8	0.97	0.97	17.0	17.6	8.9	7.0
Black spruce	Alberta	177.5	182.8	180.5	0.99	0.99	7.5	9.2	-5.4	-3.0
	Saskatchewan	181.6	178.5	180.0	0.997	0.996	5.6	6.7	3.1	1.6
	Manitoba	198.0	195.8	196.6	0.96	0.97	20.0	19.2	2.2	1.4
White spruce	Alberta	220.3	215.5	215.0	0.99	0.99	13.6	13.5	4.7	5.3
	Saskatchewan	183.1	190.1	188.4	0.98	0.97	13.8	18.6	-7.0	-5.3
	Manitoba	212.1	209.7	211.9	0.99	0.99	12.2	12.1	2.3	0.2
Balsam fir	Alberta	191.3	205.8	187.7	0.99	0.99	11.0	12.0	-14.4	3.6
	Saskatchewan	181.5	190.8	180.9	0.97	0.96	18.6	20.5	-9.3	0.6
	Manitoba	161.4	137.6	165.5	0.88	0.98	34.8	14.7	23.8	-4.1
Tamarack	Alberta	157.7	158.8	155.3	0.99	0.99	11.2	10.5	-1.1	2.4
	Saskatchewan	156.5	155.0	159.7	0.97	0.96	15.2	19.8	1.6	-3.2
	Manitoba	158.5	159.0	157.9	0.99	0.99	11.4	11.1	-0.6	0.6
Hardwoods:										
Trembling aspen	Alberta	220.2	233.6	208.2	0.99	0.97	12.6	20.5	-13.4	12.0
	Saskatchewan	195.9	189.2	203.8	0.99	0.99	7.6	9.0	6.7	-7.9
	Manitoba	212.7	206.1	216.9	0.98	0.97	13.1	18.2	6.6	-4.3
Balsam poplar	Alberta	165.5	173.8	161.8	0.97	0.97	17.7	17.8	-8.4	3.7
	Saskatchewan	141.8	142.0	146.2	0.97	0.97	16.7	18.2	-0.2	-4.4
	Manitoba	145.7	137.2	145.1	0.95	0.97	21.4	16.6	8.5	0.6
White birch	Alberta	294.9	288.0	281.6	0.98	0.99	16.1	14.0	6.9	13.4
	Saskatchewan	226.9	235.8	229.6	0.87	0.90	37.3	35.1	-8.9	-2.8
	Manitoba	234.0	231.9	244.9	0.95	0.96	22.6	21.4	2.1	-10.9

a/ Model I: $W = a_0 + a_1 D^2 H$

Model II: $W = a_0 + a_1 D + a_2 D^2 + a_3 D^3$

where W is the oven-dry weight (kg) of living tree above ground without dead branches, D is diameter (cm) outside bark at breast height, H is total tree height (m), $a_1 \dots a_3$ are regression coefficients, Residual is actual minus predicted, and % SEE is standard error of estimate expressed as percentage of the mean.

b/ Calculated from actual measurements from species sampled in Alberta, Saskatchewan, and Manitoba.

Table 2. Comparative estimates and errors in prediction when prairies general equations and individual province equations based on two regression functions are used to predict oven-dry biomass of eight species in each province^a

Species	Equation used for prediction	Alberta			Saskatchewan			Manitoba		
		Predicted Mean	R ²	% SEE	Predicted Mean	R ²	% SEE	Predicted Mean	R ²	% SEE
Model I:										
Softwoods										
Jack pine	Regional	162.94	0.97	17.2	192.48	0.98	12.5	188.95	0.97	17.8
	Provincial	157.52	0.97	16.8	189.08	0.98	12.3	197.89	0.98	15.3
Black spruce	Regional	182.83	0.99	7.2	178.51	0.997	5.7	195.81	0.96	20.2
	Provincial	177.48	0.997	5.8	181.61	0.997	5.1	197.97	0.96	19.9
White spruce	Regional	215.53	0.99	14.0	190.06	0.98	13.3	209.74	0.98	12.4
	Provincial	220.29	0.99	13.1	183.00	0.98	12.1	212.12	0.99	12.2
Balsam fir	Regional	205.77	0.99	10.2	190.83	0.97	17.7	137.55	0.88	40.9
	Provincial	191.31	0.99	7.4	181.50	0.97	16.9	161.39	0.95	23.3
Tamarack	Regional	158.76	0.99	11.1	154.96	0.97	15.3	159.03	0.99	11.3
	Provincial	157.68	0.99	10.8	156.52	0.98	14.3	158.50	0.99	11.2
Hardwoods										
Trembling aspen	Regional	233.60	0.99	11.9	189.18	0.99	7.8	206.08	0.98	13.5
	Provincial	220.20	0.99	10.7	195.90	0.997	5.3	212.63	0.98	12.7
Balsam poplar	Regional	173.84	0.97	16.9	141.97	0.97	16.7	137.16	0.95	22.7
	Provincial	165.43	0.98	16.3	141.78	0.97	16.7	145.68	0.97	17.8
White birch	Regional	288.01	0.98	16.5	235.78	0.87	35.9	231.92	0.95	22.8
	Provincial	294.93	0.98	15.6	226.90	0.88	35.2	233.99	0.95	21.8
Model II:										
Softwoods										
Jack pine	Regional	174.66	0.95	20.7	178.75	0.97	19.1	190.84	0.97	18.3
	Provincial	157.53	0.98	14.5	189.08	0.98	15.5	197.92	0.97	16.8
Black spruce	Regional	180.52	0.99	9.0	180.02	0.996	6.7	196.62	0.97	19.4
	Provincial	177.50	0.995	7.7	181.67	0.99	6.5	197.98	0.97	18.3
White spruce	Regional	215.02	0.99	13.8	188.39	0.97	18.1	211.89	0.99	12.1
	Provincial	220.27	0.99	10.5	183.14	0.97	17.7	212.10	0.99	10.6
Balsam fir	Regional	187.73	0.99	12.2	180.90	0.96	20.5	165.47	0.98	14.3
	Provincial	191.33	0.99	8.0	181.44	0.97	18.7	161.34	0.98	14.2
Tamarack	Regional	155.30	0.99	10.7	159.72	0.96	19.4	157.91	0.99	11.1
	Provincial	157.76	0.99	10.1	156.54	0.96	19.6	158.46	0.99	10.3
Hardwoods										
Trembling aspen	Regional	208.16	0.97	21.7	203.75	0.99	8.7	216.95	0.97	17.9
	Provincial	220.18	0.98	16.2	195.87	0.997	6.4	212.75	0.98	15.4
Balsam poplar	Regional	161.78	0.97	18.2	146.20	0.97	17.6	145.08	0.97	16.6
	Provincial	165.55	0.97	17.5	141.86	0.98	15.4	145.67	0.98	15.2
White birch	Regional	281.55	0.99	14.7	229.65	0.90	34.7	244.90	0.96	20.5
	Provincial	294.87	0.99	11.7	226.86	0.93	28.1	234.00	0.96	20.4

a/ Model I: $W = a_0 + a_1 D^2 H$

Model II: $W = a_0 + a_1 D + a_2 D^2 + a_3 D^3$

where W is oven-dry weight (kg) of living tree above ground (excluding dead branches), D is diameter outside bark at breast height (cm), and H is total tree height (m). Mean is predicted mean, and % SEE is standard error of estimate expressed as a percentage of predicted mean.

It is natural to expect that the individual province equations should fit the input data of the same province better than those derived elsewhere. A test was therefore made to see how good the fit for each species was when the equations derived from one province were applied to the independent data sets from the other two provinces. Table 3 shows the results of Alberta equations applied to Saskatchewan and Manitoba. In Saskatchewan, the prediction equations based on Model I showed that all species had R^2 for the estimated values of 0.97 or greater, with the exception of white birch (0.86). For Model II, all species had prediction R^2 values of 0.92 or greater, with the exception of white birch (0.88). In Manitoba, the corresponding Model I values were 0.94 or greater, with the exception of balsam fir (0.84); and the Model II prediction R^2 values were 0.92 or higher for all species. The highest fit ($R^2 = 0.99$ and % SEE = 8.3, as averaged for both models) for the use of Alberta equations in Saskatchewan and Manitoba was for black spruce in Saskatchewan.

The results of Saskatchewan equations applied to the independent data sets for Alberta and Manitoba are shown in Table 4. In Alberta, the R^2 for the estimated values ranged from 0.96 to 0.99 for all species for Model I, and from 0.93 to 0.99 with the exception of white spruce (0.87) for Model II. The corresponding values in Manitoba were as follows: Model I: 0.93 to 0.98, with the exception of balsam fir (0.85); and Model II: 0.93 to 0.99 for all species. Black spruce in Alberta showed the best predictions averaged for the two models ($R^2 = 0.99$, % SEE = 9.1).

Manitoba equations applied to prediction in other provinces showed results as listed in Table 5. In Alberta, Model I prediction R^2 values ranged from 0.93 to 0.99 for all species, with the exception of balsam fir (0.82). The Model II values ranged from 0.93 to 0.98 for all species. For Saskatchewan, Model I had R^2 values of 0.95 to 0.997 for all species, with the exception of balsam fir (0.79) and white birch (0.85). The corresponding values for Model II ranged from 0.94 to 0.99 for all species, with the exception of white birch (0.88). On the basis of the average taken for both models, trembling aspen in Saskatchewan gave the best predictions ($R^2 = 0.99$, % SEE = 9.7).

There are two obvious inferences from the error analysis presented in Tables 3 to 5: (1) with the exception of balsam fir and white birch in most cases, the equations from one province applied to another province yielded nearly as good results as the equations for the specific

province; and (2) the species with the lowest fit in all such applications over the three individual provinces was balsam fir, based on regression function of Model I. A test on the identity of individual province equations using regression functions of Models I and II showed that the Model I predictions equations were highly significantly different ($P < 0.01$) for both slope and intercept (Singh 1986). There was no such combined difference, at this probability level, for any of the remaining seven species for either model. It was therefore concluded that: (1) selection of a suitable regression function was an important consideration in deriving prediction equations, and (2) most of the separate equations could be combined into single biomass prediction equations for regional application over the Boreal Forest in western Canada.

Fresh data collection for validation purposes is a costly undertaking. A search for biomass data showed that information required for independent testing was available for two hardwoods (trembling aspen and balsam poplar), as reported by Johnstone and Peterson (1980) for six different locations in Alberta. Four of these locations were in the Boreal Forest Region and one each in the Montane Forest Region and the Forest-grassland transition (Rowe 1972). The errors involved in predicting known population parameters for the 254 trees of trembling aspen and 60 trees of balsam poplar are shown in Table 6. It is evident that the regional equations tested in the study showed better or as good a prediction for trembling aspen as the predictions obtained from the individual province (Alberta) equations. The accuracy and the errors for regional vs. Alberta equations in the case of trembling aspen were: $R^2 = 0.99$ vs. 0.97 (Model I), and 0.95 (Model I and Model II); % SEE = 15.4 vs. 22.2 (Model I), and 30.2 vs. 29.1 (Model II). For balsam poplar, the predictions for regional vs. Alberta equations were almost identical: $R^2 = 0.96$ vs. 0.96 (Model I), and 0.95 vs. 0.96 (Model II); % SEE = 23.3 vs. 24.8 (Model I), and 28.5 vs. 26.0 (Model II).

The Model I regression function gave better biomass estimates compared to the Model II regression function. For trembling aspen, Model I provided R^2 values of 0.99 vs. 0.95 and % SEE as 15.4 vs. 30.2 for regional equation, and R^2 values of 0.97 vs. 0.95 and % SEE as 22.2 vs. 29.1 for the Alberta equation. The Model I regression function also gave slightly better results for the validation data predictions of balsam poplar, mainly because the regression function of Model I is based on two measurements (D and H), rather than on a single measurement (D), as is the Model II regression function.

Table 3. Alberta equations applied to estimation of ovendry biomass of tree species in other prairie provinces^a

Species	Province	Actual Mean ^b	Predicted Mean		Predicted R ²		% SEE		Mean Residuals	
			Model I	Model II	Model I	Model II	Model I	Model II	Model I	Model II
Softwoods:										
Jack pine	Saskatchewan	189.1	185.8	163.1	0.98	0.92	12.5	29.4	3.3	26.0
	Manitoba	197.9	182.4	174.0	0.96	0.97	19.9	25.5	5.5	23.9
Black spruce	Saskatchewan	181.6	173.3	175.6	0.99	0.99	8.4	8.1	8.3	6.0
	Manitoba	198.0	190.0	194.3	0.95	0.96	20.9	20.8	8.0	3.7
White spruce	Saskatchewan	183.1	194.1	202.2	0.97	0.94	15.9	24.9	-11.0	-19.2
	Manitoba	212.1	214.3	223.8	0.99	0.97	12.7	18.1	-2.3	-11.8
Balsam fir	Saskatchewan	181.5	176.6	183.4	0.97	0.94	17.6	26.0	4.9	-1.9
	Manitoba	161.4	124.1	166.4	0.84	0.98	40.2	16.6	37.3	-5.0
Tamarack	Saskatchewan	156.5	154.0	162.5	0.97	0.96	16.2	20.3	2.6	-6.0
	Manitoba	158.5	157.9	160.7	0.99	0.99	11.2	11.8	0.5	-2.2
Hardwoods:										
Trembling aspen	Saskatchewan	195.9	176.3	214.3	0.98	0.97	12.9	18.0	19.6	-18.4
	Manitoba	212.7	193.0	231.4	0.98	0.92	16.0	30.1	19.7	-18.7
Balsam poplar	Saskatchewan	141.8	134.9	149.7	0.97	0.97	17.7	19.3	7.0	-7.9
	Manitoba	145.7	130.2	148.4	0.94	0.97	25.0	16.8	15.5	-2.7
White birch	Saskatchewan	226.9	241.0	237.0	0.86	0.88	38.8	37.9	-14.1	-10.1
	Manitoba	234.0	237.0	252.4	0.95	0.95	22.0	22.8	-3.0	-18.4

a/ Model I: $W = a_0 + a_1 D^2 H$

Model II: $W = a_0 + a_1 D + a_2 D^2 + a_3 D^3$

where W is the ovendry weight (kg) of living tree above ground without dead branches, D is diameter (cm) outside bark at breast height, H is total tree height (m), and $a_1 \dots a_3$ are regression coefficients;

% SEE is standard error of estimate expressed as percentage of the actual mean.

b/ Calculated from actual measurements from species sampled in Alberta, Saskatchewan, and Manitoba.

Table 4. Saskatchewan equations applied to estimation of ovendry biomass of tree species in other prairie provinces^a

Species	Province	Actual	Predicted Mean		Predicted R ²		% SEE		Mean Residuals	
		Mean ^b	Model I	Model II	Model I	Model II	Model I	Model II	Model I	Model II
Softwoods:										
Jack pine	Alberta	157.5	160.4	186.6	0.97	0.91	17.0	31.7	-2.9	-29.0
	Manitoba	197.9	185.7	202.4	0.97	0.97	18.6	17.3	12.2	-4.6
Black spruce	Alberta	177.5	186.0	182.8	0.99	0.99	9.2	9.0	-8.5	-5.3
	Manitoba	198.0	199.2	199.1	0.96	0.96	19.9	19.6	-1.1	-1.1
White spruce	Alberta	220.3	207.1	210.2	0.98	0.98	17.1	15.3	13.1	10.0
	Manitoba	212.1	201.6	206.1	0.98	0.98	14.2	14.0	10.4	6.0
Balsam fir	Alberta	191.3	195.7	187.8	0.99	0.97	8.7	18.4	-4.3	3.5
	Manitoba	161.4	131.0	167.7	0.85	0.98	39.4	15.6	30.4	-6.3
Tamarack	Alberta	157.7	160.5	152.1	0.98	0.99	13.7	11.2	-2.8	5.6
	Manitoba	158.5	160.8	154.7	0.98	0.99	13.4	11.7	-2.4	3.8
Hardwoods:										
Trembling aspen	Alberta	220.2	242.1	201.3	0.98	0.96	16.1	22.4	-21.9	18.9
	Manitoba	212.7	213.5	210.9	0.98	0.97	13.6	19.3	-0.8	1.8
Balsam poplar	Alberta	165.5	173.2	153.5	0.97	0.93	17.3	28.9	-7.8	11.9
	Manitoba	145.7	137.0	143.0	0.95	0.97	21.9	19.2	8.7	2.7
White birch	Alberta	294.9	273.6	243.4	0.96	0.87	22.4	41.4	21.3	51.5
	Manitoba	234.0	223.4	243.2	0.93	0.93	27.2	28.4	10.5	-9.2

a/ Model I: $W = a_0 + a_1 D^2 H$

Model II: $W = a_0 + a_1 D + a_2 D^2 + a_3 D^3$

where W is the ovendry weight (kg) of living tree above ground without dead branches, D is diameter (cm) outside bark at breast height, H is total tree height (m), and $a_1 \dots a_3$ are regression coefficients; % SEE is standard error of estimate expressed as percentage of the actual mean.

b/ Calculated from actual measurements from species sampled in Alberta, Saskatchewan, and Manitoba.

Table 5. Manitoba equations applied to estimation of oven-dry biomass of tree species in other prairie provinces^a

Species	Province	Actual Mean ^b	Predicted Mean		Predicted R ²		% SEE		Mean Residuals	
			Model I	Model II	Model I	Model II	Model I	Model II	Model I	Model II
Softwoods:										
Jack pine	Alberta	157.5	170.3	181.7	0.95	0.93	21.7	28.4	-12.8	-24.2
	Saskatchewan	189.1	201.6	185.8	0.97	0.98	16.8	16.0	-12.5	3.3
Black spruce	Alberta	177.5	184.7	182.6	0.99	0.98	9.2	14.7	-7.3	-5.2
	Saskatchewan	181.6	180.3	183.8	0.997	0.99	5.2	10.9	1.3	-2.2
White spruce	Alberta	220.3	217.9	218.2	0.99	0.98	13.6	17.1	2.4	2.1
	Saskatchewan	183.1	192.5	186.7	0.98	0.96	14.2	19.0	-9.4	-3.6
Balsam fir	Alberta	191.3	247.8	182.5	0.82	0.98	43.3	15.4	-56.4	8.8
	Saskatchewan	181.5	228.8	176.1	0.79	0.97	46.5	19.8	-47.3	5.5
Tamarack	Alberta	157.7	158.2	156.4	0.99	0.98	10.8	13.6	-0.5	1.3
	Saskatchewan	156.5	154.5	160.2	0.97	0.96	15.8	21.3	2.0	3.7
Hardwoods:										
Trembling aspen	Alberta	220.2	239.9	205.7	0.98	0.95	14.3	25.6	-19.8	14.4
	Saskatchewan	195.9	195.9	202.0	0.995	0.99	7.1	12.3	0.03	-6.2
Balsam poplar	Alberta	165.5	186.6	169.4	0.93	0.93	27.0	28.6	-21.1	-3.9
	Saskatchewan	141.8	151.0	149.2	0.95	0.94	22.2	25.8	-9.2	-7.4
White birch	Alberta	294.9	293.6	276.7	0.98	0.98	15.9	14.2	1.4	18.2
	Saskatchewan	226.9	238.1	219.4	0.85	0.88	40.0	37.4	-11.2	7.5

a/ Model I: $W = a_0 + a_1 D^2 H$

Model II: $W = a_0 + a_1 D + a_2 D^2 + a_3 D^3$

where W is the oven-dry weight (kg) of living tree above ground without dead branches, D is diameter (cm) outside bark at breast height, H is total tree height (m), and $a_1 \dots a_3$ are regression coefficients;

% SEE is standard error of estimate expressed as percentage of the actual mean.

b/ Calculated from actual measurements from species sampled in Alberta, Saskatchewan, and Manitoba.

Table 6. Comparative estimation errors for regional and individual province (Alberta) equations when validated on an independent data set to estimate oven-dry biomass of two poplar species in Alberta^a

Species	Prediction Equation	N	Actual Mean ^b	Predicted Mean		Predicted R ²		% SEE		Mean Residuals	
				Model I	Model II	Model I	Model II	Model I	Model II	Model I	Model II
Trembling aspen	Regional	254	68.0	68.6	69.8	0.99	0.95	15.4	30.2	-0.6	-1.8
	Alberta	254	68.0	57.2	69.1	0.97	0.95	22.2	29.1	10.8	-1.2
Balsam poplar	Regional	60	49.7	54.3	44.3	0.96	0.95	23.3	28.5	-4.6	5.4
	Alberta	60	49.7	50.7	47.1	0.96	0.96	24.8	26.0	-1.0	2.6

a/ Model I: $W = a_0 + a_1 D^2 H$

Model II: $W = a_0 + a_1 D + a_2 D^2 + a_3 D^3$

where W is the oven-dry weight (kg) of living tree above ground without dead branches, D is diameter (cm) outside bark at breast height, H is total tree height (m), and $a_1 \dots a_3$ are regression coefficients;

% SEE is standard error of estimate expressed as percentage of actual mean.

b/ Calculated from actual measurements from species sampled in Alberta, Saskatchewan, and Manitoba.

Conclusions

The conclusions drawn from the study are:

1. The biomass regression function based on diameter provided prediction equations with low estimation errors; the regression function based on a combination of diameter and height gave slightly better estimates.
2. The range of application of prediction equations derived from a large sample taken from widely distributed sampling locations within an ecoregion is likely to be as good or better than the sample taken from limited locations.
3. Independent data are essential for validation of a model. Validation results showed very high fit, in addition to providing more reliable estimates on the performance of regression functions tested in the study.

Literature Cited

- Johnstone, W.D.; Peterson, E.B. Above-ground component weights in Alberta *Populus* stands. Edmonton, Alberta: Environment Canada, Canadian Forestry Service, Northern Forest Research Centre; 1980; Inf. Rep. NOR-X-226. 18 p.
- Rowe, J.S. Forest regions of Canada. Ottawa, Ontario: Environment Canada, Canadian Forestry Service; 1972. Publ. 1300. 172 p.
- Singh, T. Biomass equations for ten major tree species of the prairie provinces. Edmonton, Alberta: Environment Canada, Canadian Forestry Service, Northern Forest Research Centre; 1982. Inf. Rep. NOR-X-242. 22 p.
- Singh, T. Generalizing biomass equations for the boreal forest region of west-central Canada. Forest Ecology and Management. 1986 (In press)
- Singh, T.; Campbell, D.I. FORTRAN subroutines for biomass computation. Edmonton, Alberta: Environment Canada, Canadian Forestry Service, Northern Forest Research Centre; 1983 For. Manage. Note 22. 7 p.

APPENDIX I

Biomass prediction equations for tree species in individual provinces of west-central Canada, based on 20 samples for each species per province

Species	Province	Biomass prediction equation	R ²	% SEE
Softwoods:				
Jack pine	Alberta	W ₁ = -25.75 + 6.498D - 0.1989D ² + 0.01194D ³	0.981	14.5
		W ₂ = 5.62 + 0.01749D ² H	0.971	16.8
	Saskatchewan	W ₁ = 61.22 - 15.310D + 1.2275D ² - 0.01198D ³	0.978	15.5
		W ₂ = 6.25 + 0.01775D ² H	0.985	12.3
	Manitoba	W ₁ = 14.24 - 5.986D + 0.7082D ² - 0.00371D ³	0.975	16.8
		W ₂ = 1.71 + 0.01941D ² H	0.977	15.3
Black spruce	Alberta	W ₁ = 30.18 - 10.322D + 0.9862D ² - 0.00888D ³	0.995	7.7
		W ₂ = 4.54 + 0.01810D ² H	0.997	5.8
	Saskatchewan	W ₁ = 20.84 - 5.755D + 0.6744D ² - 0.00322D ³	0.996	6.5
		W ₂ = 4.74 + 0.01897D ² H	0.997	5.1
	Manitoba	W ₁ = -23.51 + 6.655D - 0.1590D ² + 0.01201D ³	0.969	18.3
		W ₂ = 2.33 + 0.01909D ² H	0.958	19.9
White spruce	Alberta	W ₁ = 47.21 - 13.851D + 1.1396D ² - 0.01035D ³	0.992	10.5
		W ₂ = 7.61 + 0.01656D ² H	0.987	13.1
	Saskatchewan	W ₁ = -1.07 - 0.633D + 0.3165D ² + 0.00245D ³	0.969	17.7
		W ₂ = 11.39 + 0.01524D ² H	0.984	12.1
	Manitoba	W ₁ = 5.56 - 1.944D + 0.3273D ² + 0.00366D ³	0.991	10.6
		W ₂ = 11.76 + 0.01605D ² H	0.986	12.2
Balsam fir	Alberta	W ₁ = -15.64 + 3.657D - 0.0632D ² + 0.01011D ³	0.995	8.0
		W ₂ = 2.07 + 0.01532D ² H	0.995	7.4
	Saskatchewan	W ₁ = 18.23 - 7.804D + 0.8077D ² - 0.00689D ³	0.970	18.7
		W ₂ = 13.59 + 0.01474D ² H	0.972	16.9
	Manitoba	W ₁ = 11.76 - 4.565D + 0.5177D ² - 0.00118D ³	0.983	14.1
		W ₂ = 4.42 + 0.01970D ² H	0.947	23.3
Tamarack	Alberta	W ₁ = 20.22 - 6.754D + 0.7244D ² - 0.00601D ³	0.991	10.1
		W ₂ = 8.89 + 0.01784D ² H	0.988	10.8
	Saskatchewan	W ₁ = 23.93 - 7.219D + 0.7110D ² - 0.00556D ³	0.962	19.6
		W ₂ = -0.93 + 0.01936D ² H	0.977	14.3
	Manitoba	W ₁ = -18.18 + 3.723D + 0.0287D ² + 0.00649D ³	0.989	10.3
		W ₂ = 8.28 + 0.01798D ² H	0.985	11.2
Hardwoods:				
Trembling aspen	Alberta	W ₁ = 24.26 - 6.433D + 0.5800D ² + 0.00247D ³	0.981	16.2
		W ₂ = -9.11 + 0.01909D ² H	0.990	10.7
	Saskatchewan	W ₁ = 5.02 - 2.278D + 0.3691D ² + 0.00411D ³	0.997	6.4
		W ₂ = 0.55 + 0.02011D ² H	0.997	5.3
	Manitoba	W ₁ = 11.32 - 5.253D + 0.7389D ² - 0.00479D ³	0.980	15.4
		W ₂ = 9.44 + 0.01919D ² H	0.985	12.7
Balsam poplar	Alberta	W ₁ = 0.54 - 1.921D + 0.4346D ² - 0.00200D ³	0.975	17.5
		W ₂ = 10.38 + 0.01324D ² H	0.975	16.3
	Saskatchewan	W ₁ = 31.85 - 10.205D + 0.9447D ² - 0.01114D ³	0.979	15.4
		W ₂ = 13.74 + 0.01362D ² H	0.972	16.7
	Manitoba	W ₁ = -5.16 + 1.235D + 0.1149D ² + 0.00513D ³	0.979	15.2
		W ₂ = 6.27 + 0.01540D ² H	0.968	17.8
White birch	Alberta	W ₁ = 24.48 - 7.202D + 0.7906D ² - 0.00160D ³	0.990	11.7
		W ₂ = 0.53 + 0.02602D ² H	0.979	15.6
	Saskatchewan	W ₁ = 110.55 - 39.002D + 3.1177D ² - 0.04645D ³	0.934	28.0
		W ₂ = 18.50 + 0.02255D ² H	0.883	35.2
	Manitoba	W ₁ = 13.04 - 5.423D + 0.6400D ² + 0.00075D ³	0.964	20.4
		W ₂ = -9.31 + 0.02677D ² H	0.954	21.8

a/ Model I: $W_1 = a_0 + a_1 D^2 H$

Model II: $a_0 + a_1 D + a_2 D^2 + a_3 D^3$

where W is the oven-dry weight (kg) of living tree above ground (excluding dead branches), D is the diameter outside bark at breast height (cm), and H is total tree height (m).

APPENDIX II

Prairies regional equations for biomass prediction derived from 60 trees per species in the three provinces

Species	Biomass Prediction Equation ^a	R ²	% SEE
Softwoods:			
Jack pine	$W_1 = 4.09 + 0.01829D^2H$	0.975	15.3
	$W_2 = 13.87 - 4.998D + 0.5954D^2 - 0.00159D^3$	0.967	17.9
Black spruce	$W_1 = 4.06 + 0.01871D^2H$	0.982	12.9
	$W_2 = 6.70 - 3.040D + 0.5131D^2 - 0.00047D^3$	0.984	12.4
White spruce	$W_1 = 8.88 + 0.01609D^2H$	0.984	12.8
	$W_2 = 10.39 - 3.749D + 0.4906D^2 + 0.00018D^3$	0.983	13.5
Balsam fir	$W_1 = 13.57 + 0.01556D^2H$	0.952	21.5
	$W_2 = 8.15 - 3.034D + 0.4074D^2 + 0.00112D^3$	0.978	14.9
Tamarack	$W_1 = 5.89 + 0.01833D^2H$	0.982	12.2
	$W_2 = 16.08 - 5.135D + 0.5945D^2 - 0.00358D^3$	0.980	13.3
Hardwoods:			
Trembling aspen	$W_1 = 1.41 + 0.01933D^2H$	0.988	11.2
	$W_2 = 21.73 - 7.304D + 0.7545D^2 - 0.00307D^3$	0.977	15.8
Balsam poplar	$W_1 = 12.23 + 0.01380D^2H$	0.966	18.0
	$W_2 = 6.54 - 3.432D + 0.5021D^2 - 0.00295D^3$	0.974	16.3
White birch	$W_1 = 2.89 + 0.02520D^2H$	0.943	24.4
	$W_2 = 30.26 - 10.591D + 1.0458D^2 - 0.00686D^3$	0.956	21.8

a/ Model I: $W_1 = a_0 + a_1 D^2H$

Model II: $W_2 = a_0 + a_1 D + a_2 D^2 + a_3 D^3$

where W is oven-dry weight (kg) of living tree above ground (excluding dead branches), D is the diameter outside bark at breast height (cm), and H is total tree height (m).

FOREST BIOMASS STUDIES

IN FRANCE

Daniel AUCLAIR

INRA, Centre de Recherches d'Orléans
Station de Sylviculture
ARDON - 45160 OLIVET (France)

Several research institutes or universities, as well as forest managers, are interested in biomass production and estimation. This paper discusses research under way in the field of biomass production, in traditional forestry (high forest, coppice with standards, simple coppice) as well as in "future" forestry, such as short rotation coppice or "modified conventional forestry".

Introduction

Although some forest biomass studies had been initiated in the late sixties (RIEDACKER, 1968) they never lead to real research programs in forest institutes before the creation of government agencies interested in the use of biomass for energy production. First the COMES (Commissariat à l'Energie Solaire), "solar energy agency", which became later AFME (Agence Française pour la Maîtrise de l'énergie), "french energy agency", produced financial incentives for research oriented towards biomass production. The European Economic Community later became interested in biomass production and launched several european research programs.

In France, the global afforested area represents approximately one-fourth of the country : 14 million hectares. One-third of the forest consists of conifers, and two-thirds of broad-leaves, of which 2.5 million hectares are treated as simple coppice and 4.25 million hectares as coppice-with-standards.

The ownership of French forests is divided into 1.7 M ha state owned, 2.5 M ha managed by the state, and 10.8 privately owned, by one and a half million owners. Private property is very diverse, and private forests larger than 100 ha only amount to 2 M ha. (AUCLAIR, 1982).

Research is undertaken at different levels by basic and applied institutes : universities, CNRS (Centre National de la Recherche Scientifique) : the "scientific research institute", INRA (Institut National de la Recherche Agronomique) : "agricultural research institute", AFOCEL (Association Forêt-Cellulose) : "pulp and paper association", CEMAGREF (Centre d'Etude du Machinisme Agricole, du Génie Rural et des Eaux et Forêts) for machinery and applied research.

Two main research areas concern forest biomass : existing forest stands, and short rotation forestry (TEISSIER DU CROS, 1985).

1 - Forest biomass in existing stands

Forest stands under simple coppice or coppice-with-standards management have not been harvested regularly since the 1940's. These stands produce mostly poor quality timber, and have a low volume (or biomass) production. The first studies have concerned biomass inventories.

One government institution : IFN (Inventaire Forestier National) has the charge of the French forest inventory, estimating for the total French forest area standing volume, current annual increment, number of trees, and area occupied by each forest species, for each forest region, and for different ownerships.

This comprehensive survey gives estimates of "large timber volume" (LTV), which only concerns stem volume down to a diameter of 7 cm. Several studies have been undertaken to convert LTV into total aboveground woody biomass in these poor quality stands as well as in Oak high forest. Simple weighted linear regressions were applied on a selection of sample trees in Central France, giving the following results (AUCLAIR and BIGE, 1984 ; AUCLAIR and CABANETTES, 1983 ; BISCH, 1985).

- for coppiced trees, from simple coppice or from coppice-with-standards :

$$B = 0.569 \text{ LTV} + 7.9$$

- for oak standards from coppice-with-standards :

$$B = 1.011 \text{ LTV} - 74.0$$

- for oak from traditional high forest :

$$B = 0.729 \text{ LTV} - 3.3$$

where B = total above-ground woody biomass (kg)
LTV = "large timber volume" (dm³)

The distribution of biomass inside each individual tree has been studied, to compare the effect of high-forest or coppice-with-standards treatment on the larger oak trees. The results in figure 1 show that in high-forest, trees are higher and have smaller crowns (BISCH and AUCLAIR, 1987).

Coppice, which produces mostly poor quality wood which may be used for industry or energy, has been quite extensively studied, and some biomass production curves are now being published for Sweet Chestnut or Black Locust as in figure 2 (PAGES, 1986).

The above equations have also been used to convert traditional high forest volume yield tables (PARDE, 1962) into biomass yield tables (BISCH, 1987).

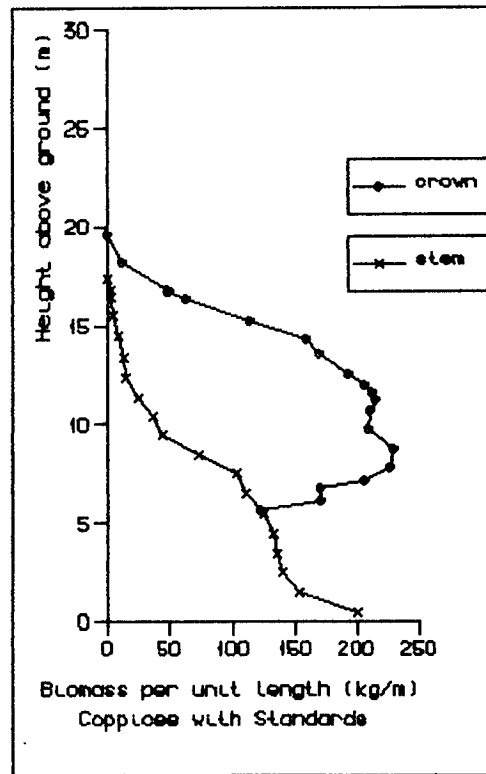
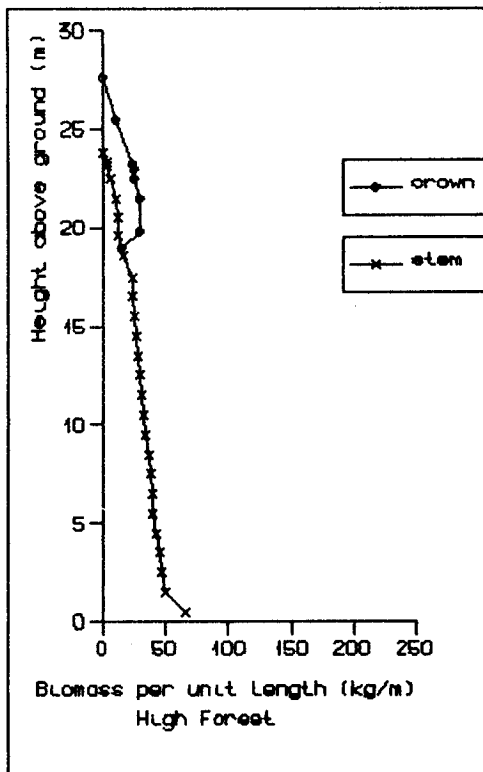


Figure 1 : Biomass profiles for Oak

High Forest
age 122
DBH 34 cm
Biomass 841 kg

Coppice-with-standards
121
60 cm
2 873 kg

BLACK LOCUST CURRENT INCREMENT

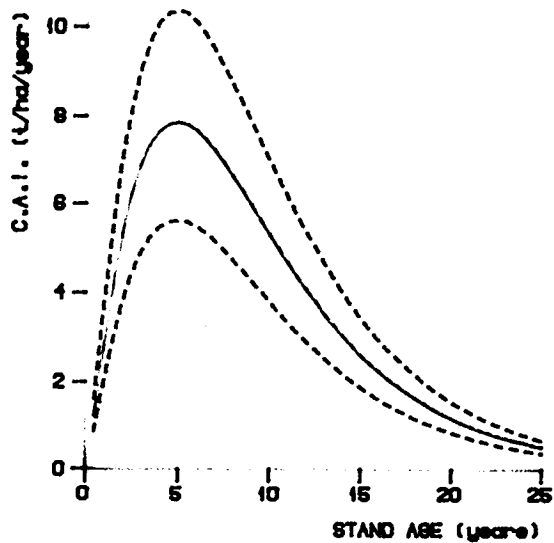


Figure 2 : Standing biomass production curves for traditional Robinia coppice in Central France.

- solid line : mean for the region,
- dashed lines : maximum and minimum values

2 - Short rotation forestry

There is a wide potential for new forest plantations on marginal forest or agricultural sites. One important program concerns genetic selection and improvement for short-term biomass production, with quite a large number of tree species : poplars, larch, Sitka spruce, red oak, yellow poplar, sugi, alder, black locust.

Silvicultural aspects concern establishment, cultivation, and harvesting of short rotation forestry, including both coppice or short rotation single-stem biomass production.

Competition aspects of short rotation coppice have involved the use of "Nelder fans" : figure 3 shows growth curves of one-year rotation black cottonwood coppice. It shows a decrease in production from one rotation to the next, more important for closer than for wider spacings. This decrease was not observed for two-year rotations (AUCLAIR, 1986 a).

POPULUS 1-YEAR ROTATION

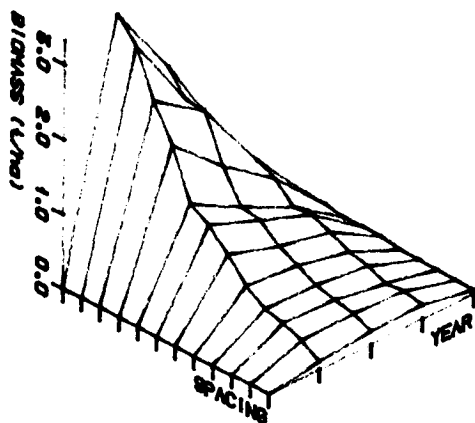


Figure 3 : Short rotation *Populus trichocarpa* annual production, from a Nelder design. Spacings vary from 0.15 m to 1.5 m.

Some economic studies have been undertaken to compare various degrees of intensification in biomass production, from a traditional "low-cost : low-return" extensive coppice forestry to short rotation intensive coppice production. Up to now most economic studies concerning biomass have focused on energy production, and it seems that the present current wood price cannot compete with oil if intensive cultivation is practised on forest land (AUCLAIR, 1985).

Attention is now turning to agricultural land for biomass production, with a new type of "modified conventional forestry" : some trials have been established with genetically selected

improved trees for multiple use (biomass for industry or energy, and high quality lumber), with modern highly mechanized techniques.

Most research on short rotation forestry for biomass production, or for multiple use, is only quite recent, and most experimental plots have not yet given many results, in particular those concerning coppice. Some plots are still being established.

3 - The scientific background

Although most french studies concerning the physiology and growth of forest trees are oriented towards high quality lumber production, several laboratories have been interested in forest biomass production. Most of these studies concern coppice, either treated in short rotation with the aim of producing biomass, or traditional coppice which can have several different uses. This second type, treated in 20- to 30-year rotations can contribute precious information for biomass-oriented short rotation coppice.

A working party gathering research scientists and forest professionals has been meeting every year since 1982 to discuss the scientific basis of coppice production. The topics studied concern :

- bud formation
- early growth and shoot demography
- gas exchange
- carbohydrate reserves
- nutrient cycling
- nitrogen fixation
- root systems
- fine root production
- root-shoot interactions
- pests and diseases

A summary of the results is presented in papers by AUCLAIR (1986 a, 1986 b).

Conclusions

Since the first research studies impulsed by government incentives in the late seventies, a number of publications have concerned traditional forestry, mostly with the use of results of the French national forest survey. Some trials have been established in existing stands to test techniques aimed at increasing biomass production and the impact of these techniques on the nutrient balance.

Studies on short rotation forestry, concerning genetic selection and improvement, silvicultural techniques and environmental problems (including nutrient cycling, pests and diseases), have lead to the establishment of over 50 hectares of trials oriented towards basic research, as well as an additional 240 hectares oriented towards development.

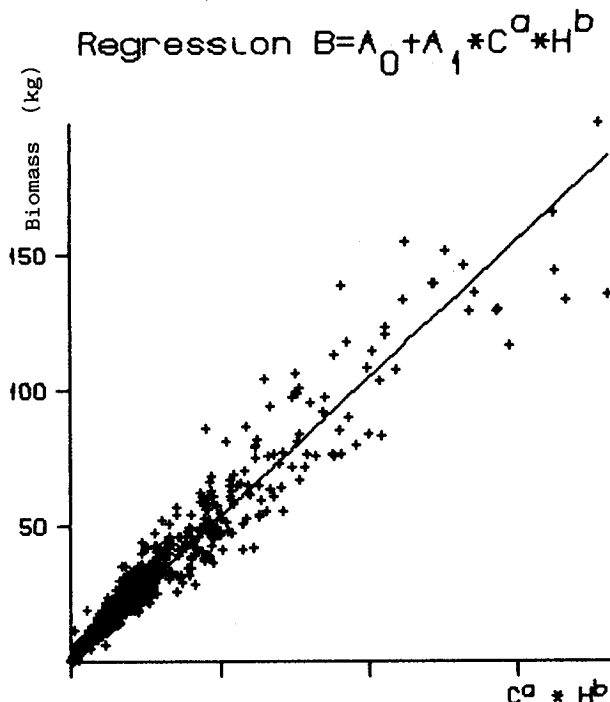
All these biomass studies have lead to the construction of a data base for improving biomass regressions. Non-linear weighted regressions have been used, giving quite satisfactory results (figure 4) with an equation of the following form :

$$B = a_0 + a_1 D^\alpha H_d^\beta$$

where B = biomass

D = diameter at breast height

H_d = stand dominant height



Literature cited

AUCLAIR, D. - 1982 : Present and future management of coppice in France. Forestry commission and Institute of chartered foresters Symposium : Broadleaves in Britain, future management and research. Loughborough (GBR). Malcolm, D. C. ; Evans, J. ; Edwards, P. N. (Ed.). 1982/07/07-09. Forestry Commission, Farnham (GBR), 40-46.

AUCLAIR, D. - 1985 : Fast growing trees as an alternative use of farmland. The French research program and some implications. Workshop on "Agricultural surpluses", group "lignocellulosic biomass and processing". Oberursel (DEU) 1985/12/9-10. European federation of biotechnology, 174-182.

AUCLAIR, D. - 1986 (a) : The scientific background. In : Hummel, F. C. (Ed.) - Forest biomass plantations in the E. C. Commission of European Communities, 9 p.

AUCLAIR, D. - 1986 (b) : Coppice versus single-stem : physiology, growth, economics. 18th IUFRO world congress. Ljubljana (YU). 1986/09/07-21, 14 p.

AUCLAIR, D. ; BIGE, M. C. - 1984 : Une méthode d'évaluation régionale de la biomasse des taillis à partir des données de l'Inventaire Forestier National. Application à la Région Centre. Ann. Sci. for., 41 (4), 405-426.

AUCLAIR, D. ; CABANETTES, A. - 1983 : Method for the estimation of biomass in coppice and coppice-with-standards, by conversion from "large timber" volume to total biomass. Energy from biomass, 2nd E. C. conference. Berlin (DEU). Strub, A. ; Chartier, P. ; SCHLESER, G. (Ed.). 1982/09/20-23. Appl. Sci. Publ., London, 222-224.

BISCH, J. L. - 1985 : Influence du traitement sylvicole (futaie et taillis sous futaie) sur la répartition de la biomasse dans le chêne en Région Centre. INRA, Station de Sylviculture, Orléans, 85/42, 26 p.

BISCH, J. L. - 1987 : Un exemple de conversion d'une table de production en volume en tables de production en biomasse : le Chêne dans le secteur ligérien. Ann. Sci. for., in press.

BISCH, J. L. ; AUCLAIR, D. - 1987 : Influence of the silvicultural treatment (coppice with standards, high forest) on oak aboveground biomass distribution in central France. Forestry, in preparation.

PAGES, L. - 1986 : Lois de croissance en biomasse du taillis : le robinier dans le Val de Loire. Ann. Sci. for., 43 (4), 533-550.

PARDE, J. - 1962 : Table de production pour les forêts de Chêne rouvre de qualité tranchage dans le secteur ligérien. Notes Tech. for. 11, Station de recherches et expériences forestières, Nancy, 6 p.

RIEDACKER, A. - 1968 : Méthodes d'estimation de la biomasse d'un arbre. Dissertation, D.E.A. écologie, Université Paris Sud (Orsay), 31 p.

TEISSIER DU CROS, E. - 1985 : Forest biomass. INRA's program. Energy from biomass, 3rd E.C. conference. Venice (ITA). Palz, W. ; Coombs, J. ; Hall, D.O. (Ed.). 1985/03/25-29. Appl. Sci. Publ., London, 305-309.

BIOMASS STUDIES IN EUROPE--AN OVERVIEW

Dieter R. Pelz

Professor of Forest Biometry, Universität
Freiburg, 7800 Freiburg i. Br., W-Germany

Research on estimating biomass components of forests have long tradition, in the last two decades considerable progress has been made to provide information on the biomass of entire stands. Research projects can be defined in three major classes (1) biological studies, (2) wood utilization studies (whole tree utilization, weight scaling) and (3) biomass energy. Information on many species for different sites and stand structures is available but few reliable data on the biomass potential of large regions or countries exist, showing the need for including biomass information in national inventories or conducting special biomass inventories.

Biomass studies have gained worldwide increasing attention and support. Within the last few years the literature on such studies has increased tremendously, each year many new projects are reported and new data published.

Several bibliographic studies were made for a review of the literature, such as Keays (1971) who reviewed estimation procedures and results for the main components of tree biomass separately. Hitchcock & McDonnell (1979) cite 280 references, and Pardé (1981) reviewed biomass research in a forestry Abstract review Article. Cannell reports a total of about 300 projects of which 131 are European studies, another review was done by Cupi (1982).

During recent years many symposia and workshops were conducted, the corresponding proceedings contain a good selection of study reports. The set of volumes put out by the IUFRO working group that was initiated by Harold Young should be mentioned especially, the IUFRO biomass studies 1973, Oslo biomass studies 1976. In addition the proceedings from the biomass conference in Orleans, France, in 1983 contain a set of papers dealing with biomass studies in Europe, a special conference of studies in Scandinavia was conducted in 1979 in Umea, Sweden, the papers are presented in Albrektson et al (1980). Egnéus and Ellegård present papers of a conference in 1984 conducted in Sweden that deal also with forest biomass.

In the following presentation an overview of European forest biomass studies will be given, however I cannot claim to provide a complete coverage, due to the large number of published studies and due to the diversity of sources (in several languages), often in form of local research reports. Some important and valuable publications will have been missed for

which I apologize to the authors, they can be sure that such omissions are unintentional. No Russian references are being considered as literature in Russian is difficult to gain access to, however in Forestry Abstracts many reviews are listed.

This overview is mainly concerned with mensurational aspects, and not with problems of wood utilization or biological considerations. Foresters are traditionally concerned with measuring trees and forests to determine their potential use. Over time one has to consider various utilization standards, and various wood uses. Estimating the availability of fuel wood was quite important for a long time in European forest history (in some parts this is gaining again in importance). Volume estimates therefore often included information on branch and twig volume or percent.

One of the first foresters to deal with what we consider now biomass (or a part thereof) was the German forester Hartig, who in 1896 studied the relation of the weight of branches incl. needles to wood increment on the basis of 5 spruce trees in Würzburg, and stressed the importance of weight measurements.

Tischendorf from Austria described in 1922 how to measure the various tree biomass components solid wood, branch and twig volume, and bark. The yield tables of Grundner-Schwappach 1922 include information of branches and twigs in percent of stem volume for age classes, diameter, and height classes. Measurement of root and stock volume was made by Prebler and Kunze, and Vanselow (1941) gives a detailed treatment of measuring tree components by volume and weight, incl. specific gravity, weight increment and others.

For two sites the figures were calculated for model trees, the data for the Wiener Wald site are shown in table 1. Vanselow states that a ranking of volume yield or increment is not equal to a ranking of weight yield or increment, thus recognizing the importance of weight studies. He discusses the importance of measuring the entire organic substance (i.e. biomass) incl. wood, leaves, needles, seed, roots. A detailed discussion of weight measures and specific gravity determination is given by Trendelenburg 1939 and other books on forest utilization.

The most detailed studies on components of forest biomass were done by Burger in Switzerland between 1922 and 1953. These studies were for many years the main data base for a large number of other projects and were cited quite often. Burger studied both single trees and stands for many tree species, different ages and site classes. A summary of the data on leaf mass and leaf surface is given by Pelz (1969); Pardé (1980) lists data on the density of wood that was reported by Burger, many other authors referred to these studies.

TABLE 1 Fresh weight of leaves and surface areas of several species for dbh classes (Burger 1922).

	diameter class (cm)									
	10		20		30		40		50	
	kg	m ²	kg	m ²	kg	m ²	kg	m ²	kg	m ²
Spruce selection forest	5,7	29,6	21,8	113,4	48,0	249,6	85,5	444,6	132,4	688,5
Fir selection forest	6,9	42,1	31,2	190,3	65,9	402,0	111,8	682,0	163,0	994,0
Fir even age forest	6,6	40,0	23,9	133,8	52,6	294,6	88,0	492,8	129,0	722,4
Pine	4,4	24,2	11,6	63,8	21,1	116,1	34,5	189,8	51,7	284,4
Larch	2,4		5,4		10,4		17,5		28,0	
Beech	3,1	62,0	9,1	182,0	18,7	374,0	34,3	686,0	54,0	1080,0
Oak	2,7	35,1	7,8	101,4	17,7	230,1	33,0	429,0	52,0	676,0
Douglas Fir	5,2	33,0	6,2	39,0	20,0	126,0	50,2	321,0	-	-

TABLE 2 Weight and increment of trees from Wiener Wald (Vanselow, 1939)

	Beech	Oak	Spruce	Fir	Pine	Larch
Volume (m ³) age 110	0,3454	0,6529	1,6524	1,479	1,5643	1,5544
Spec. gravity kg/cm ³	570	570	390	370	420	490
Total weight of wood	197	372	644	547	657	386
Average increment kg	2,2	4,1	7,1	6,1	7,3	8,5

Table 2 gives a summary of leaf mass and surface for several tree species for dbh classes based on the studies of Burger.

In Denmark Boysen-Jensen (1932) and Møller (1945) reported on studies and results related to biomass.

Tree properties, wood density, specific gravity on one side and biological aspects on the other were studied in many projects in most European countries for many years, rather independent and without references to each other.

Major impetus for biomass research came from several sources:

- growing environmental awareness called for more and detailed information on the natural resources.
- utilization standards and new harvesting methods (weight scaling, complete tree harvesting), changed and required studies on the entire tree.
- energy shortage - the energy crisis in the early 70's lead to a large number of projects dealing with forest biomass as energy source.

The projects of the first type, shortly called biological projects, were conducted in most countries in Europe, the studies were rather varied in objective, methodology and reported results, so that they really are not comparable. With the initiation of the International Biological Program 1963, the number of research studies in this field increased even further, but much better coordinated, directly in form of projects in the IBP program and in form of other studies that were modeled after these projects. In the following some of the projects will be discussed.

In the IBP program a total of 31 research sites in Europe (table 3) were included, constituting the woodlands data set. The detailed analyses, study setup and a listing of the entire data set is given by Reichle (1981), Cannel (1982) published a set of data showing the results of biomass studies from a large number of international projects.

An analysis of these results shows the diversity of these data, they hardly can be used to draw any definite conclusions on the availability of biomass resources. It is remarkable that a large number of the studies dealt with young stands, sampling procedures and measurement techniques differed considerable.

Table 3. IBP study sites in Europe.

Country	Number of study sites
Belgium	1
CSR	1
Denmark	1
Finland	1
France	3
Hungary	1
Netherlands	1
Poland	3
Rumania	4
Spain	1
Sweden	6
United Kingdom	1
West-Germany	7

Individual tree biomass in most cases was determined by weighing the components, sampling procedures were used rather infrequently. Schöpfer (1961) for example used a multistage sampling procedure to estimate total crown and needle mass. The studies determined either fresh weight or oven dry weight. Total biomass for a stand was determined either by the mean tree method by diameter classes or by regression estimates.

The mean tree method as used by Wright and Will (1958), Ovington (1957), and others selects trees of average size (mostly the tree average basal area). A number of trees with diameters close to the diameter of the tree is then chosen, felled and the total biomass measured. The actual number of trees measured varies from 1 to about 10, for reliable estimate at least 5-10 should be chosen (Madgwick and Satoo 1982). This mean tree method is consistent with methods used in the estimation of stand volumes (however for volume (biomass) estimation the tree of average volume (biomass), approximated by the Weise tree - 40% of the distribution from the upper end-might be more appropriate).

The accuracy of this estimate is difficult to ascertain, as few complete biomass inventories of entire stands exist, but it must be noted that in many cases too few trees were measured to provide sufficient precision. This mean tree method gives only an estimate of the total biomass of the stand, and no distribution over size classes.

Ovington and Madgwick (1959), Vyskot (1983), Mounet (1978) and others determine diameter classes for the stand and determine the mean tree biomass for each diameter class. The total biomass then is calculated by multiplying mean tree values by the number of trees per hectare in each diameter class. Vyskot (1972) divided the stand into 3 classes and sampled 5 dominant, 5 codominant and 5 subdominant trees for biomass. Total stand biomass then was

calculated by multiplying the respective biomass by the number of trees in that class.

Attiwill and Ovington (1968) compare alternative methods of deriving stand estimates by the mean tree method, by taking 4 observations in each of 5 diameter classes and by taking the average biomass from regression estimates, they find deviations about -3% for method 2, -8% for method 1 and -11% for method 3.

The use of regression models for estimating the total biomass per hectare is reported by many authors. Albrektson (1980), Decei (1981) and others use a model with diameter as independent variable, Anderson (1981), Ellenberg (1981) a.o. us D^2H , Mounet (1978) and Ranger (1978) use the circumference.

Models used for deriving these regression equations ranged from linear model of the type

$$\begin{aligned} B &= b_0 + b_1 D + b_2 H \\ B &= b_0 + b_1 D^2 H \end{aligned} \quad \begin{aligned} \text{where } B &= \text{biomass} \\ D &= \text{DBH} \\ H &= \text{height} \end{aligned}$$

to nonlinear models (Attiwill and Ovington (1968))

$$\begin{aligned} \text{where} \quad DW &= b_0 + b_1 D \\ DW &= \text{dry weight} \end{aligned}$$

$$\begin{aligned} \text{The model} \quad y &= b_0 + b_1 X \\ \text{where} \quad y &= \text{biomass} \\ X &= \text{diameter or circumference at} \\ &\quad \text{breast height} \end{aligned}$$

has also been used extensively.

The majority of studies seem to develop nonlinear models, the allometric equations have been introduced into ecological research long time ago and are well established. The bias inherent in this estimation is not in all cases recognized and corrected. The linear models generally are not appropriate as one of the basic assumptions of the OLS estimate, the homogeneity of variance, is not fulfilled.

The use of weighted least squares analyses for biomass studies as reported in the literature is not widespread.

With respect to biomass functions and tables one notes a large number of equations, all of which cannot be listed here. With these tables such as developed by Hellrigl (1974) in Italy for silverfir or Ranger (1978) in France for corsican pine, Kestemont (1975) in Belgium for several stands, a large number of tables are published in Sweden as part of a biomass table project, one can predict the biomass production of stands, some even can be used to construct biomass yield tables. Other tables were reported by Vyskot (1981) and others.

The data base for these biomass yield tables in many cases is much smaller than the data base for standard yield tables, however they give some insight in the biomass production over time. Estimation of stock and root biomass is reported from various silvicultural or ecological studies on the root system of tree species (Köstler 1950) and from the special biomass studies (Vyskot 1972 and others). The measurement problems here are quite large, and reliable data difficult to obtain. As Pardé (1980) points out only the larger roots are of practical importance for possible utilization, making the estimation procedure somewhat more easy to implement. However if ecological objectives are most important then also the small roots have to be considered which are less than 2 mm (data of this kind are not as complete as it should be, but many studies were done as reported by Hermann in a review of the literature). The root biomass can be related to DBH, diameter at stumpheight, or the total height, in addition population and site variables will influence root biomass.

Several attempts have been made to derive biomass information from stand timber inventory estimates. Some of these conversions are based on detailed studies such as the equations developed by Auclair (1984) in France, who sampled 96 stands of coppice to derive an equation of the form

$$BST = b_0 + b_1 VBF$$

I.N.R.A. projects developed methods to determine the aboveground biomass in coppice stands allowing a reliable estimate. Ruprich (1981) discusses the possibilities of converting inventory data with equations taken from special studies and the literature, in part going back to the data reported by Burger (1922-53). For entire Tschechoslowakia the LESPROJECT estimated the total biomass by converting inventory data. Similar estimations are reported by Dauber and Kreutzer (1979) and Kramer and Krüger (1981) for Germany.

Some of the studies cited above were made to estimate the energy potential of forest, special studies on energy plantation are numerous, however they will not be reviewed here, as the mensurational aspects are not of significance there. There are many studies on forest biomass from the standpoint of forest utilization, the dry weight for the various components are listed, specific gravity and complete tree utilization discussed.

Gislerud (1974), Hakkila (1975), Eskilson (1972) and Palenius (1974) studied specific gravity of the various tree components separate for diameter classes.

Wernius (1975) listed tables giving the percentages of dry weight for the biomass components of spruce and pine, Hakkila (1977) developed regression equations for biomass for

bole, branches and the complete tree. Schäfer (1977) studied biomass production in young pine stands in the southern Germany to predict potential utilization and May (1976) made a similar study in the Lüneburger Heide.

Most of these studies used similar measurement techniques and calculation procedures as mentioned above, therefore they will not be reviewed in detail here.

After the energy crisis 1973 an increasing amount of attention has been given to the use of forest biomass as source of energy, many research projects were initiated after that time. For example the European common market initiated biomass projects in 1975, in three phases. The first phase, that ran to 1985 intended to utilize side products from forestry and agriculture, without change of the present land use, in the second phase energy plantations were planned and in the third phase these plantations would be increased considerably if deemed necessary.

Within that program several projects were initiated, the majority with the objective of determining energy potential of trees and forests. Another project that can be listed here is the Swedish Energy Forest Project. With the increasing concern about nuclear energy it can be expected that research on biomass from agriculture and forestry will become even more important in the future.

The plans for utilization of the forest resource at high intensities resulted in concerns that the site might be degraded over time as much more of the production is taken out and less is recycled into the soil.

Krapfenbauer (1981) for example conducted in Austria an inventory of biomass and nutrient elements in 4 spruce stands at the ages of 10, 45, 52 and 63. For these stands he developed linear regression models for DBH, height, volume, and biomass for several tree components and the distribution of the nutrient elements.

The general agreement among many forest nutritionists seems to indicate that on extreme sites with complete utilization one has to monitor the site quality carefully and in some cases has to fertilize or restrict the utilization, but that, taken over an entire rotation the additional removal on many sites does not result in site degradation as long as not the entire annual biomass production (leaves etc.) is harvested.

Information on biomass for many trees and stands in Europe is available, the large number of studies made during the last few years is quite impressive the bibliography, which certainly is not exhaustive comprises more than 180 entries. Large scale inventories of forest biomass on a regional or national level have to

be conducted to yield reliable estimates on the total forest resources.

The data on biomass production reported in the literature are not comparable, as they were derived by different procedures, many of the studies were very detailed and provided information on the reliability of the estimates, others were designed to yield approximate estimates only. For the major species one can find information on biomass production at various ages, and in some cases for different sites. This information can be quite valuable for estimating biomass for the given study site, however it does not suffice for the estimation for a larger area, for example during a forest inventory. Factors for converting traditional forest inventory estimates, mostly relating to wood production, to biomass estimates can be derived from these studies, as several reports show, the extrapolation to other data, however, can yield only rough estimates or educated guesses.

It would seem necessary to conduct such inventories, either in combination with national forest inventories (which are done in most European countries) or separately, comprehensive inventory models need to be developed to accurately predict biomass for entire countries of entire Europe as basis for an efficient use of the natural resources.

Literature Cited

- Albrektson, A. 1981. Biomass of Scots pine. Amount--development--methods of mensuration. Rapport, Institutionen for Skogsskotsel, Sveriges Lantbruksinstitut 2, 189 p.
- Albrektson, A.; Lindroth, S.; Stömne, P.E. (Hrsg). 1980. Biomass research in the Nordic countries. A symposium in Umea 23-25 Oct. 1979. Rapport, Institutionen for Skogstaxering, Sveriges Lantbruksuniversitet 29, 175 p.
- Albrektson, A.; Valinger, E.; Jonsson, C. 1984. New biomass functions for parameters of Scots pine in southern Norrland. Sveriges Skogvårdsförbunds Tidskrift 82 (6), 5-12.
- Alestalo, A. 1973. On the possibilities of the utilization of needles and bark. IUFRO Biomass Studies, University of Maine, Orono. 429-442.
- Amri, A. 1979. Verwendung der Biomasse der Baumarten Schwarzkiefer (*Pinus nigra*) und Robinie (*Robinia pseudoacacia* L.) als Spanplattenrohstoff unter Berücksichtigung der Ganzbaumernte im Kurzumtrieb. Doctoral thesis, Univ. Freiburg. 143 p.
- Andersson. 1981. In: Dynamic properties of forest ecosystem. Reichle ed., Oxford Univ. Press. 620 p.

- Andersson, F. 1970. Ecological studies in a Scanian woodland and meadow area. South. Swed. II. Plant biomass, primary production and turnover of organic matter. *Botaniska Notiser* 123, 8-15.
- Attiwill, P.M.; Orington, J.D. 1968. Determination of forest biomass. *For. Sci.* 14, 13-15.
- Auclair, D. 1983. Les methodes de mesure de la biomasse forestiere. *Biomasse Actualities* 19, 14-15.
- Auclair, D. 1985. Short rotation coppice forest biomass production. The work of IUFRO S1.05-10 Working Party. Energy from biomass, 3rd E.C. conference, 25-29.3 1985. Venice. 274-277 *Appl. Sci. Publ.*, London.
- Auclair, D.; Bige, M.C. 1984. Une methode d'évaluation régionale de la biomasse des taillis a partir des données de l'inventaire forestiere national. Application a la region Centre. *Annales des Sciences Forestieres* 41 (4), 405-425.
- Auclair, D.; Cabanettes, A. 1983. Method for the estimation of biomass in coppice and coppice-with-standards, by conversion from "large timber" volume to total biomass. Energy from biomass, 2nd E.C. conference 20-23.9 1982. Berlin, 222-224, *Appl. Sci. Publ.*, London.
- Auclair, D.; Jorez, J.P.; Piermont, L. 1985. Variabilite des qualites energetiques de plaquettes du Bouleau en Sologne. *Rev. for. Fr.* 37 (5), 389-397.
- Auclair, D.; Maerten, E. 1986. Un methode d'evaluation de la biomasse des arbres de halle. *Ann. Sci. for.* 43 (1), 59-67.
- Auclair, D.; Metayer, S. 1980. Methodologie de l'evaluation de la biomasse aerienne sur pied et de la production en biomasse des taillis. *Acta oecol., Oecol. appl.* 1 (4), 357-377.
- Aussenac, G. 1969. Litter production in four young Douglas fir stands in eastern France. *Revue For. Franc.* 31 (1), 15-19 - CNRF, Champenoux, Seichamps, Frankreich
- Bedeneau, M.; Pages, L. 1984. Etude des carnes d'accroissement ligneux du systeme racinaire d'arbres traites en taillis. *Ann Sci. for.* 41 (4), 59-68.
- Belyakov, P. 1979. The aerial biomass of red oak. *Gorskostopanska Nauka* 16 (3), 12-26 - Inst. za Gorata, Sofia, Bulgarien.
- Bouchon, J. 1973. Biomass studies in a biological reserve I. Estimation of branch volume. *IUFRO Biomass Studies*, 251-257 - Coll. of Life Sciences and Agriculture, University of Maine, Orono.
- Bouchon, J.; Nys, C.; Ranger, J. 1985. Cubage, biomasse et mineralomasse: comparaison de trois taillis simples des Ardenes primaires. *Acta Oecologica, Oecologia Plantarum* 6 (1), 53-72.
- Bouchon, J.; Riedacker, A. 1973. Allometric relationship in a young coppice of *Eucalyptus camaldulensis*, Indications for the above-ground biomass sampling. *IUFRO Biomass Studies*, 259-268 - Coll. of Life Sci. and Agric., Univ. of Maine, Orono.
- Boysen, Jensen P. 1932. Die Stoffproduktion der Pflanzen. 108 p. - G. Fischer, Jena.
- Brandles, L. et al. 1984. Energy forestry aspects in general. In: Ecology and management of forest biomass production systems. Rapport, Institutionen for Ekologi och Miljövard - Sveriges Lantbruksuniversitet.
- Brunig, F.F. 1973. Biomass Diversity and Biomass Sampling in Tropical Rainforest. *IUFRO Biomass Studies*, 269-294 - University of Maine, Orono.
- Burger, H. 1929. Holz, Blattmenge und Zuwachs 1. Mitteilung: die Weymouthsföhre. *Mttlg. der Schweizerischen Anstalt für das Forstl. Versuchswesen* 15, 243-292.
- Burger, H. 1935. 2. Mitteilung: die Douglasie. *ibid.* 19, 21-72.
- Burger, H. 1937. 3. Mitteilung: Nadelmenge und Zuwachs bei Föhre und Fichte. *ibid.* 20, 101-114.
- Burger, H. 1940. 4. Mitteilung: ein 80jähriger Buchenbestand. *ibid.* 21, 307-348.
- Burger, H. 1941. 5. Mitteilung: Fichten und Föhren verschiedener Herkunft auf verschiedenen Kulturorten. *ibid.* 22, 10-62.
- Burger H. 1942. 6. Mitteilung: ein Plenterwald mittlerer Standortsgüte. *ibid.* 22, 377-445.
- Burger H. 1945. 7. Mitteilung: die Lärche. *ibid.* 24, 7-103.
- Burger, H. 1947. 8. Mitteilung: die Eiche. *ibid.* 25, 211-279.
- Burger, H. 1948. 9. Mitteilung: die Föhre. *ibid.* 25, 435-493.
- Burger, H. 1950. 10. Mitteilung: die Buche. *ibid.* 26, 419-468.

- Burger, H. 1951. 11. Mitteilung: die Tanne. *ibid.* 27, 247-286.
- Burger, H. 1952. 12. Mitteilung: Fichten in Plenterwald. *ibid.* 28, 109-156.
- Burger, H. 1953. 13. Mitteilung: Fichten im gleichaltrigen Hochwald. *ibid.* 29, 38-130.
- Cabanettes, A.; Rapp, M. 1981. Biomasse, mineralomasse et productivité d'un écosystème à pins pignous (*Pinus pinea* L.) du littoral méditerranéen IV. Production. *Acta Oecologica, Oecologica Plantarum* 2 (4), 381-394.
- Cannell, M.G.R. (compiler). 1982. World forest biomass and primary production data. Academic Press, London.
- Cantiani, M. 1974. Tavola di produttività della biomassa arborea. Ricerche sperimentali. V. Prime indagini sulla biomassa, 41-57 - Istituto di Assest. For., Fac di Agr., Univ. Firenze.
- Carey, M.L. 1980. Whole tree harvesting in Sitka spruce, Possibilities and implantations. *Irish Forestry* 37 (1), 48-63.
- Carre, J. 1978. The use of forest biomass for manufacture of particleboard. *Bulletin des Recherches Agronomiques de Gembloux* 13 (4), 287-309 - Station de Technologie Forestiere.
- Chen, T.Y.; Paulitsch, M. 1974. Eigenschaften und Verwendungsmöglichkeiten von Biomasse-Waldhackschnitzeln. *Mitt. FVA Freiburg* H62, 69-94.
- Clabault, G.; Lemée, G. 1980. Investigation on the ecosystems of the nature reserve of the forest of Fontainbleau 7, Structure and functions of holly (*Ilex aquifolium*) stands. *Revue d'Ecologie* 34, 317-334 - Lab. Phytotron, CNRS, Gif-sur-Yvette, Frankreich.
- Cupi, F. 1982. Forest biomass. Bibliographic review 1950-1981. *Institute National de la Recherches Forestieres, Nancy, France.*
- Dauber, E.; Kreutzer, K. 1979. Die Ermittlung des Potentials forstlicher Reststoffe in der Bundesrepublik Deutschland. *FoWI* CB1. 98 (6), 289-297.
- Decei, I. 1981. Biomass of High Productivity Trees and Young Beech Stands (*Fagus silvatica* L.). *Kyoto Biomass Studies*, 125-128 - University of Maine, Orono.
- Devillez, F.; Gossuin, A.; Marynen, T. 1973. Productivité écologique et rendement photosynthétique d'une hêtraie en Haute-Ardenne. *Bulletin, Classe de Sciences, Acad. Royal de Belgique* 59, 431-440.
- Dietz, P. 1966. Gewichtsvermessung von Industrieholz - warum? *Allg. Forstzeitung* Wien 58, 81-84.
- Dietz, P.; Jehl, R. et al. 1966. Vermessung von Industrie-Rohholz nach Gewicht. *Holzzentralblatt* 92, 1233-1236.
- Dimitrov, E.P. 1983. Studies on the above-ground phytomass in high-yield spruce stands. *Gorskostopanska Nauka* 20 (5), 45-51.
- Dimitrov, E.P. 1984. A weight yield table for spruce stands of high site class. *Gorskostopanska Nauka* 21 (3), 37-44.
- Dimitrov, E.T. 1982. Analysis of the suitability of some regression models for expressing crown biomass in Scots pine. *Gorskostopanska Nauka* 19 (3), 33-41.
- Dissescu, G. & R. 1983. Relations dimensionnelles nécessaires pour la détermination expéditive de la superficie foliaire du *Quercus pedunculiflora* Koch. *Les Colloques de l'INRA* 19, 311-319.
- Dissescu, R. 1973. Contributions to the Study of the Correlation between the Size of Foliage and Current Volume Growth for Trees of *Robinia pseudacacia* L. *IUFRO Biomass Studies*, 295-304 - University of Maine, Orono.
- Dissescu, R.; Tabrea, A.; Cecei, I. 1981. A Complete Biomass Study in Young Stands of *Abies alba* Mill. *Kyoto Biomass Studies*, 129-138 - University of Maine, Orono.
- Droste zu Hulshoff, B. von. 1968. Vorläufige Untersuchungsergebnisse über die Erfassung oberirdischer Baumorgane an einer 76 jährigen vorh. Fichte im Ebersberger Forst bei München. *FwCbl* 6, 321-384.
- Duvigneaud, P. (Hrsg.) 1971. Productivity of forest ecosystems. *Proceedings of the Brussels symposium organized by Unesco and the International Biological Programme*, 707 p. - UNESCO, Paris.
- Duvigneaud, P.; Kestemont, P. 1977. Productivité biologique en Belgique. 617 p.
- Ekersten, H.; Nilsson, L.O. 1983. Radiation and temperature influences on deciduous forest production. *Les Colloques de l'INRA* 19, 153-165.
- Egnéus, H.; Ellegård, A. ed. 1984. Biomass resources: Bioenergy 84. *Proc. of conf.* 15-21.6.1984. Elsevier Appl. Sci. Publ.
- Elias, P. 1983. Biomass estimation of *Loranthus europaeus* coenopopulation in an Oak-Hornbeam forest. *Les Colloques de l'INRA* 19, 321-322.

- Ellenberg. 1981. In: Dynamic properties of forest ecosystems. Reichle ed., 669-671 - Oxford Univ. Press.
- Elowson, S.; Rytter, L.; Ericsson, T.; Kormo, H. 1984. Distribution of biomass and species. In: Ecology and management of forest biomass production systems. Rapport, Institutionen for Ekologi och Miljövara - Sveriges Lantbruksuniversitet.
- Erickson, H. et al. 1984. Analyses of growth and yields. In: Ecology and management of forest biomass production systems. Rapport, Institutionen for Ekologi och Miljövard - Sveriges Lantbruksuniversitet.
- Eskilsson. 1972. Whole tree pulping, Part 1 Fibre properties. Svensk Papperstidn. 75, 397-402.
- Faber, P. J.; van den Burg, J. 1982. The production of woody biomass. In Forest and Energy, Nederlands Bosbouw tijdschrift 54 (7/8), 198-205.
- Ford, E.D. 1982. High productivity in a potstage Sitka spruce stand and its relation to canopy structure. Forestry 55 (1), 1-17.
- Frivold, L.H.; Borchgrevink, I. 1981. Biomass yield of *Betula pendula* in a 6-yr-old trial plantation at Ås, Norway. Meldinger, Norges Landbrukshogskole 60 (12), 17 p.
- Galoux, A. 1981. Total production of wood. Bulletin de la Société Royale For de Belgique 88 (3), 85-96.
- Galvaus, U.I. 1976. Die oberirdische Biomasse bei der Durchforstung gefällter Bäume. IUFRO Congress Oslo.
- Garelkov, D. 1973. Biological Productivity of some Beech Forest Types in Bulgaria. IUFRO Biomass Studies, 305-314 - University of Maine, Orono.
- Gislerud, O. 1974. Whole tree utilization II, Biomass and biomass properties of trees from thinnings of spruce, pine, birch and alder. Norsk Institutt for Skogforskning.
- Grundner-Schwappach. 1922. Massentafel zur Bestimmung des Holzgehaltes stehender Waldbäume und Bestände. 6.Aufl. - Herausg. R. Schober, Berlin+.
- Gulbrandsen, K. 1977. Whole tree utilization - future quantities from first thinnings. Meddelelser fra Norsk Institutt for Skogforskning 33 (12), 481-512 - Norw. For. Res. Inst., As.
- Hägg, S. 1975. Determination of root biomass of spruce sample trees. Rapporter och Uppsatser, Inst. for Västekologie och Marklara 19, 37 p. - Royal College of Forestry, Stockholm.
- Hakkila P. 1967. Variation patterns of bark percentage by weight. Communicationes instituti forestalis Fenniae 62, 5.
- Hakkila P. 1979. Wood density survey and dry weight tables for pine, spruce and birch stems in Finland. Communicationes instituti forestalis Fenniae.
- Hakkila, P. 1969. Weight and composition of the branches of large Scots pine and Norway spruce trees. Communicationes Instituti Forestalis Fenniae 67 (6), 37 p.
- Hakkila, P. 1970. Basic density, bark percent and dry matter content of grey alder (*Alnus incana*). Finnish Forest Research Institute.
- Hakkila, P. 1971a. Coniferous branches as a raw material source. Finnish Forest Research Institute.
- Hakkila, P. 1971b. Branches, stumps and roots as a future raw material source in Finland IUFRO Working Group on Forest Biomass Studies. Univ. of Florida, Gainesville; Univ. of Maine, Live Sciences and Agric. Exp. Station 1971, 51-62.
- Hakkila, P. 1972. Mechanized harvesting of stumps and roots. Finnish Forest Research Institute.
- Hakkila, P. 1973a. Towards the Utilization of Stump and Root Wood. IUFRO Biomass Studies, 451-466 - University of Maine, Orono.
- Hakkila, P.; Makela, M. 1973. Harvesting of stump and root wood by the Palari Stumparvester. Finnish Forest Research Institute.
- Hartig, R. 1896. Wachstumsuntersuchungen an Fichten. Forstl. Naturwiss. Zeitschrift Bd.1+2.
- Heinonen, J.; Ferm, A. 1983. Aboveground woody biomass sampling and estimation in dense birch coppice stands. Les Colloques de l'INRA 19, 209-216.
- Hellrigl, B. 1974. Relazioni e tavole della biomassa arborea. Inst. Asses. Forest. Univ. Firenze - Ric.speriment Dendro.Auxoment 5, 1-40.
- Hitchcock, H.C., III; McDonnell, J.O. 1979. Biomass measurement: a synthesis of the literature. 59 p. Proc.-Division of Land and Forest Resources, Norris, Tennessee.

- Jozefacinkowa, W. 1975. The biomass of the root system of the pine and oak in the Kampinos National Park. *Ekologia Polska* 23, 83-92.
- Kanninen, K.; Uusvaara, O.; Valonen, P. 1979. Measurement and properties of the whole-tree raw material. *Folia Forestalia, Inst. For. Fenniae* 403, 53 p. - For. Res. Inst., Unioninkatu, Helsinki.
- Kanninen, M.; Hari, P.; Kellomäki, S. 1982. A dynamic model for above-ground growth and dry matter production in a forest community. *Journal of Applied Ecology* 19 (2), 465-476 - Dept. Silviculture, Univ. Helsinki.
- Kestemont, P. 1972. Biomasse, productivité et phytogéochimie de la végétation riveraine d'un ruisseau ardennais. *Bull. Soc. roy. Bot. Belg.* 105, 209-219.
- Kibzun, A.P.; LeChong Kuk. 1979. Attempt at mathematical evaluation of the vertical fractional distribution of the phytomass of the aerial part of a forest-steppe oak stand. *Lesovedenie* 5, 19-25 - Inst. Geog., MGU, Moskau.
- Kleinschmit, J.; Knigge, W. 1967. Durch Umwelt und Erbanlagen bedingte Variation der Trockensubstanzerzeugung, Struktur und Rohdichte junger Fichten (*Picea abies* Karst). *AFZ* 138 (9), 189-198.
- Köstler. 1950. Waldbau - Grundriss und Einführung, 418 p. - Parey, Berlin.
- Kommert, R. 1982. Mitteilungen über Holzeigenschaften der Serbischen Fichte (*Picea omorica*) und der Stechfichte (*P. pungens* Engelm). *Holztechnologie* 23 (2), 115-116.
- Kramer, H. 1983. Möglichkeiten und Grenzen einer Biomassennutzung im Wald. *AFZ* 38/23, 582-585.
- Kramer, H.; Krüger, H.H. 1981. Vorrat und Nutzungsmöglichkeiten forstlicher Biomasse in der BRD. *FoHo* 36/2, 33-37.
- Krapfenbauer, A. 1981. Holzernte-Biomassen- und Nährstoffaustrag - Nährstoffbilanz eines Fichtenbestandes. *Intern. Holzmarkt Jg. 72*, 25/26, S.10-13.
- Krapfenbauer, A.; Buchleitner, E. 1981. Holzernte, Biomasse- und Nährstoffaustrag - Nährstoffbilanz eines Fichtenbestandes. *Centralblatt ges. Forstw. Jg. 98*, 4, pp 193-223.
- Krapfenbauer, A. 1979. Waldbiomasse und Energieversorgung. *Int. Holzmarkt* 4/5, 26-32.
- Krapfenbauer, A. 1980. Holzernte - Biomassen - und Nährstoffaustrag. reprint.
- Kreutzer, K. 1979. Ökologische Fragen zur Vollbaumernte. *FoWi CB1*. 98(6), 298-308.
- Kreutzer, K.; Dauber, E. 1980. Die potentielle forstliche Nutzung in der BRD. *Forstw. CB1*. 99/1, 1-5.
- Kruger, J. 1984. Untersuchung zur Ermittlung von Baummassen. *Wiss. Zeitschrift der TU Dresden*.
- Kuusela, K.; Nyyssönen, A. 1981. The forest biomass as an energy source: the role of the forest inventory. *Tiedonantoja, Helsingin Yliopisto, Metsänarvioimistieteen Laitos* 14, 11 p.
- Lamotte, M.; Bourlière, F. 1978. Problèmes d'écologie: structure et fonctionnement des écosystèmes terrestres. 345 pp.
- Lemoine, B.; Gelpe, J.; Ranger, Jr.; Nys, C. 1983. Biomasses et minéralomasses du Pin Maritime Etude d'un peuplement de 16 ans. *Les Colloques de l'INRA* 19, 189-198.
- Mälkönen, E. 1973. Effect of Complete Tree Utilization on the Nutrient Reserves of Forest Soils. *IUFRO Biomass Studies*, 375-386 - University of Maine, Orono.
- Mälkönen, E.; Saarsalmi, A. 1982. Biomass production and nutrient removal in whole tree harvesting of birch stands. *Folia Forestalia, Inst. Forest. Fenniae* 534, 17 p.
- Marchenko, A.; Rok'yanis, B. 1978. Biological productivity of the *Piceetum hylocomiosum* and *Piceetum oxalidosum*. *Trudy LSKhA (LLA Raksti)* 143, 56-65 - LLA Elagava, Lett. SSR.
- Marklund, L.G. 1983. Collecting data for biomass equation development: some methodological aspects. *Les Colloques de l'INRA* 19, 37-43.
- May, H.M. 1976. Waldhackschnitzel für die Spanplattenherstellung. *Holz-Zentralblatt*, 569 p.
- Merzlenko, M.D.; Gurtsev, A.I. 1982. Biological productivity of Scots pine plantations in relation to planting density. *Lesovedenie* 2, 85-88.
- Mieghem, A. van; Schalck, J.; Stevens, M. 1983. Productivity of single stem poplar plantations in Belgium. *Les Colloques de l'INRA* 19, 179-188.
- Miller, H.G.; Miller, J.D.; Cooper, J.M. 1980a. Biomass and nutrient accumulation at different growth rates in thinned plantations of Corsican pine. *Forestry* 53 (1), 23-39.

- Miller, H.G.; Miller, J.D.; Cooper, J.M. 1980b. Tables of biomass and accumulated nutrients at different growth rates in thinned plantations of Corsican pine. 31 p. - Macaulay Inst. for Soil Res., Aberdeen.
- Mitchell, C.P.; Proe, M.F.; MacBrayne, C.G. 1981. Biomass tables for young conifer stands in Britain. Kyoto Biomass Studies, 45-50.
- Moller, C. 1945. Untersuchungen über Laubmenge, Stoffverlust und Stoffproduktion des Waldes. Dff Danm. 17, 1-287.
- Moskalayuk, T.A. 1980. The amount and structure of the phytomass of stands in the maintypes of larch forests in the northern part of the Sea of Okhotsk region. Lesovedenie 2, 32-39 - Magadan. Lesn. Opyt. St., Dal'NIILKh, UdSSR.
- Mounet, J.P. 1978. Evaluation de la biomasse aérienne des chenes pubescents. Ph.D. thesis.
- Neenan, M.; Steinbeck, K. 1979. Caloric values for young sprouts of nine hardwood species. Forest Science 25 (3), 455-461 - An Foras Taluntais, Oak Park, Carlow, Irland.
- Neuwirth, G. 1972. Bestandesstruktur, Stoffproduktion und Stoffbilanz eines 35 jährigen Kiefernbestandes. Arch. Natur. Landschaft 12, 101-120.
- Newbould, P.J. 1967a. Methods for Estimating the Primary Production of Forests. JBP Handbook 2 - Blackwell Scientific Publications, Oxford, Edinburgh.
- Newbould, P.J. 1967b. Methods of estimating the primary root production of forests. Blackwell Scientific Publications, Oxford.
- Nihlgard, B. 1972. Plant biomass, primary production, and distribution of chemical elements in a beech and planted spruce forest in South Sweden. Oikos 23, 69-81.
- Nilsson, L.O. 1981. Methods of determining dry matter content and growth in energy forest plantations. Tech. rep. 19, 47 p. - University of Agricultural Sciences, Uppsala.
- Nilsson, L.O. 1982. Determination of current forest growth and biomass production. Tech. Rep. 27, 36 p. - University of Agriculture Sciences, Uppsala.
- Nilsson, L.O. 1983. The preservation of leaf samples between sampling and measurement. Tech. Rep. 31, 21 p. - University of Agricultural Sciences, Uppsala.
- Orington, J.D.; Magdwich, H.A.I. 1959. The growth and composition of natural stands of birch. I- Dry matter production. Pl. Soil 10, 271-283.
- Palenius, I. 1974. Marginal raw materials in pulp manufacture. Papperi ja puu 3, 146-152.
- Pardé, J. 1980. Forest biomass. For. Abstracts Review Article 41 (8), 343-362.
- Patronov, D. 1980. Biological productivity of young stands of Quercus petraea and Q. frainetto on various sites. Gorskostopanska Nauka 17 (1), 10-20 - Opitna Stantsiya po D"b, Gori, Burgas, Bulgarien.
- Pavlov, M.B. 1972. Bioelement-Inventur von Buchen und Fichtenbeständen im Solling. Doctoral thesis, Univ. Göttingen, W.Germany.
- Pellinen, P. 1984. Notwendigkeit und Probleme der Biomassenermittlung. AFJZ 155/6, 141-143.
- Pelz, D. 1969. Blattfrischgewicht und Blattoberfläche verschiedener Baumarten und Bestandesformen (unpublished manuscript). Abt. Forstliche Biometrie, Univ. Freiburg.
- Pereira, H.; Sardinka, R. 1984. Utilization of Eucalyptus globulus Lab. for energy and fibre. Biomass production and nutrient removal assessment. Bioenergy 84 Proc. Conf. 15-21 June 1984.
- Perttu, K.L. (Hrsg.) 1984. Ecology and management of forest biomass production systems. Rapport, Institutionen for Ekologi och Miljövard 15 - Sveriges Lantbruksuniversitet.
- Pismerov, A.V.; Pismerova, R.S.; Vorobei, P.M. 1979. Phytomass productivity of young conifer and conifer/broadleaved stands. Lesovedenie 1, 68-72.
- Polák, J. 1979. Stanoreni biomasy na celkovi teszbe roční. Lesprojekt Brandys.
- Polge, H. 1966. Etablissement des courbes de variation de la densité du bois par exploration densitométrique d'échantillons prélevés a la tarière sur des arbres vivants. Annales de Sciences For. 23 (1), 215 p.
- Pollanschütz, J. 1983. Economic necessities and ecological limits of forest biomass utilization. Les Colloques de l'INRA 19, 15-20.
- Prebier, M.; Kunze, M. 1873. Die Holzmesskunde in ihrem ganzen Umfang+.
- Ranger, J. 1978. Recherches sur les biomasses comparées de deux plantations de pin laricio de Corse avec ou sans fertilisation. Annales des Sci. For. 35 (2), 93-115.
- Rapp, M.; Cabanettes, A. 1981. Biomass and productivity of a Pinus pinea L. stand. In: Margeris, N.S.; Mooney, H.A. (Ed.) - Components of productivity of mediterranean-

- climate regions. Basic and applied aspects. 131-134 - W. Junk, La Haye.
- Reichle, D.E. 1981. Dynamic properties of forest ecosystems. International Biological Programme 23 - Cambridge University Press.
- Riedacker, A. 1968. Methodes indirectes d'estimation des arbres et des peuplements forestiers. Institut National de la Recherche Agronomique, Nancy.
- Riedacker, A.; Knockaert, Z.; Zaidi, A. 1985. Production ligneuse des taillis d'eucalyptus: effect de la frequence de coupe et de la densité de plantation. Ann. Sci. For. 42, 39-52.
- Rodin, L.E.; Bazilevich, N.I. 1967. Production and mineral cycling in terrestrial vegetation. Oliver and Boyd, London.
- Rok"yanis, B. 1978. The phytomass and amount of chemical elements in spruce stands in Lativa. Trudy LSKhA (LLA Raksti) 143, 43-56 - LLA, Elgava, Lettische SSR.
- Rollinson, T.J.D. 1983. Biomass estimation for open-grown trees. Les Colloques de l'INRA 19, 31-36.
- Rossi Marcelli, A.; Duranti, G.; Giulimandi, G. 1980. The productivity of a poplar plantation II. Net primary productivity. Pubblicazioni del Centro di Sperimentazioni Agr. e For. 13 (3), 203-227.
- Ruprich, J. 1981. Möglichkeiten der Erfassung der potentiellen und realisierbaren Biomasse in der Forsteinrichtung. IUFRO Proc. 17th World Congress, 348-361.
- Saarsalmi, A.; Palmgren, K. 1984. Biomass production and nutrient requirement of a young green alder plantation. Bioenergy 84 Proc. Conf. 15-21 June 1984.
- Satoo, T.; Magdwich, H.A.I. 1982. Forest biomass, 152 p.
- Schäfer, G. 1977. Nährelementhaushalt von Kiefern-Jungbeständen in der südlichen Oberrheinebene - Doctoral thesis.
- Schöpfer, W. 1961. Beiträge zur Erfassung des Assimilationsapparates der Fichte. Schriftenreihe Ba-Wü FVA Abt. Biometrie 10.
- Schulze, E.D.; Fuchs, M.I. & M. 1977. Spacial distribution of photosynthetic capacity and performance. I. Biomass distribution. Oecologia 29 (1), 43-61.
- Scott, K.G. et al. 1982. Productivity of coppice willow in biomass trials in the U.K. Long Ashton Research Station, Bristol.
- Simola, P. 1977. The biomass of small-sized hardwood trees. Folia Forestalia, Institutum Forestale Fenniae 302, 16 p.
- Smolyak, L.P.; Rusalenko, A.I.; Petrov, E.G. 1977. Tables of the aerial phytomass of pine forests in Belorussia. Lesnoe Khozyaistvo 2, 68-71.
- Solymos, R. 1973. Untersuchungen über die Biomasse in Beständen der Gemeinen und der Schwarzen Kiefer. Erdeszeti Kulatasok 69, 181-193.
- Solymos, R. 1983. La Mesure et l'utilisation de la biomasse en Hongrie. Les Colloques de l'INRA 19, 237-242.
- Spank, G. 1982. Schätzung der Kronen- und Nadelmasse in Reinbeständen der Baumart Kiefer. Beiträge für die Forstwirtschaft 16 (3), 129-139.
- Stoiculescu, S. 1981. Biomass Estimation of Bald Cypress Trees in Romanian Forest Cultures. Kyoto Biomass Studies, 139-150. University of Maine, Orono.
- Stoll, H.P. 1975. Das Gewicht von Ganzbaumen bei Fichte. AFZ 30, 644 p.
- Thill, A. 1983. Biomasse ligneuse et équivalent calorique de deux taillis de 20 et 40 ans. Les Colloques de l'INRA 19, 199-207.
- Tischendorf, W. 1927. Lehrbuch der Holzmassenermittlung. 218 p.
- Trendelenburg. 1939. Das Holz als Rohstoff. Seine Entstehung, stoffliche Beschaffenheit und chemische Verwertung. 435 p. - Lehmanns, Munchen.
- Usol'tsev, V.A.; Krichun, V.M. 1982. Features in the formation of the aerial phytomass of birch and aspen in forest outliers in northern Kazakhstan. Lesovedenie 3, 41-52.
- Usol'tsev, V.A.; Makarenko, A.A.; Atkin, A.S. 1979. Features of the formation of the aerial phytomass of Scots pine in northern Kazakhstan in relation to stand density. Lesovedenie 5, 3-12.
- Uspenkii, V.V. 1982. Tables for calculating the amount of needle-bearing twigs in fellings and sample plots of Scots pine. Lesnoe Khozyaistvo 9, 55-57 - VLTII, Voronezh.
- Utkin, A.I.; Gul'be, Ya.; Ermolova, L.S. 1980. Primary production of Alnus inc. stands in the Yaroslavl region. Lesovedenie 3, 69-80 - Lab. Les., AN SSSR.
- Vanselow, K. 1941. Einführung in die forstliche Zuwachs- und Ertragslehre. 155 p. - Sauerländer, Frankfurt.

- Vins, B.; Sika, A. 1977. Biomass of the above-ground and underground parts of sample trees of Norway spruce. *Prace Vyzkumného Ustavu Leshního Hospodarství a Myslivosti* 51, 125-150.
- Vyskot, M. 1972. Aerial biomass of silver fir (*Abies alba* Mill.) *Acta Universitatis Agriculturae (Brno)* C 41, 243-294.
- Vyskot, M. 1976a. Tree story biomass in lowland forests in South Moravia. *Rozpravy Českoslov. Akademie Ved, Rada Matemat. a Přírod. Ved* 86 (10), 167 p.
- Vyskot, M. 1976b. Biomass production of the tree layer in a floodplain forest near Lednice. *Oslo Biomass Studies - 16th International Congress of IUFRO*, 117-202.
- Vyskot, M. 1976c. Floodplain Forest in Biomass. *Oslo Biomass Studies - 16th International Congress of IUFRO*, 206-229.
- Vyskot, M. 1981. *Larix decidua* Mill. in Biomass. *Kyoto Biomass Studies*, 151-162 - University of Maine, Orono.
- Vyskot, M. 1983a. Young Scotch pine in biomass. 148 p. - *Academia Nakladatelství Českosl. Akademie Ved, Prag*.
- Weber, M. 1983. Inventurverfahren zur Biomassenermittlung von Waldbeständen - Thesis. *Abt. Forstliche Biometrie, Univ. Freiburg*.
- Wernius, S. 1974. Compilation of literature -whole tree utilization. *Rap. Nr. 77*, 45 p. - *Skogshögskolan, Inst. f. Skogsteknik, Garpenberg*.
- Wilhelmsen, G.; Vestjordet, E. 1973. Preliminary Dry Wood Weight Tables for Merchantable Stems and Stands of Norway Spruce (*Picea abies* Karst.) in Norway. *IUFRO Biomass Studies*, 517-531 - University of Maine, Orono.
- Windhorst, H.W. 1979. Neuere Versuche zur Bestimmung der Primärproduktion der Walder und forstlicher Ertragspotentiale. *Erdkunde* 33 (1), 10-23.
- Wright, T.W.; Will, G.M. 1958. The nutrient content of Scots and Corsican pines growing on sand dunes. *Forestry* 31, 13-25.

SUBSAMPLING TREES FOR BIOMASS

C. Kleinn and D. R. Pelz

Graduate Research Assistant and Professor of
Forest Biometry, University of Freiburg, Germany

A sampling procedure for estimating total forest weight of coniferous trees is presented. The weight of the bole including bark is estimated by simple ratio estimates of volume/weight on the basis of five discs that are selected along the stem with a probability proportional to estimated volume. For the estimation of crown weight branches are selected with a probability proportional to branch dimension. Regression estimates for branch weight were derived using weighted least squares, the estimate of the total tree weight yielded for the 16 trees that were included in the study a small negative bias.

The estimation of the biomass production of forest ecosystems is becoming increasingly important. Forest inventory estimates are often required to include information on biomass. For some regions biomass information is available in biomass regression equations or tables, however, these estimates may or may not be applicable for a given case.

The fresh weight of an individual tree may be determined by weighing all components or by sampling. For large trees weighing of the entire tree may become quite time consuming and laborious. To minimize field work sampling procedures may be applied.

In this paper a sampling method is presented that allows the estimation of the fresh weight of coniferous trees. This study is part of several pilot studies to identify optimal sampling procedures for estimating above ground biomass in the school forest of the University of Freiburg/Germany.

The study was conducted in a 20 year old spruce stand in the school forest. The stand was homogeneous, the crowns relatively uniform developed. Sixteen sample trees were selected in the diameter range from 11 to 16 cm. These sample trees were felled by the same procedures, care was taken to minimize crown damage. For all trees the above ground biomass (stem, branches, twigs and needles) was determined in two steps. First, the green (fresh) weight of the entire tree was determined by measuring volume of the stem and total weight of the all components, and secondly, a sample of the main components was taken to calculate conversion factors from volume to weight and fresh to dry weight, respectively (for this, the samples were dried at 103° for 72 hours).

This paper will only refer to the estimational fresh weight, subsampling for oven dry weight is commonly done in most biomass studies. The main purpose of the study was not to determine overall biomass of the stand or to develop tables, but to provide information as pilot study with regard to some sampling alternatives in this area. Other studies, on a smaller scale, compared additional methods (Weber 1983). The estimation of total above ground biomass is separated into two parts, estimation of stem biomass and of crown biomass.

Estimating Stem Biomass

The procedure for estimating the stem biomass was the same for both methods. The process was defined in two stages. In the first stage the volume of the stem to a minimum diameter of 3 cm was calculated by sections. This volume measurement had to be converted to weight by determining the relative density or by taking a sample for which the conversion factor could be directly measured as ratio volume to fresh weight. The second alternative was chosen, at five locations along the stem wood discs were taken, the location was selected by a PPS procedure. Random numbers were drawn, and compared with the cumulative volume at this point, the probability of any one location being selected was proportional to the volume, the location along the stem for taking samples was calculated in the field with a HP 33 hand held calculator. At these 5 points, discs were taken for which the conversion factors were determined.

The results (Table 1) show very little deviation from the true weight, the results vary between 90% and 110% of true weight.

Table 1.--Accuracy of bole weight estimation

Tree No.	Bole Weight (kg)	
	True	Estimated
1	127.9	113.8
2	73.1	65.6
3	63.5	59.7
4	114.3	113.4
5	72.3	67.8
6	128.9	125.4
7	154.2	159.5
8	67.2	66.7
9	120.4	129.0
10	63.9	64.7
11	51.5	49.0
12	85.2	82.2
13	84.5	78.4
14	132.8	126.7
15	76.4	78.2
16	136.2	125.8
Total	1552.3	1505.9

Sampling Procedure for Crown Weight

Crown weight was determined by two methods, a cross section method and by a PPS procedure. The cross section method, also used for estimating logging residue, determined the volume of branch and twig piles, by using a ratio estimate, the total weight can be determined. In this study the results, however, were severely biased. Therefore, only the second method was developed further.

The biomass of a crown of a coniferous species can be determined by sampling branches and estimating the weight by regression of branch diameter and length to weight. Several studies have found that this relationship is sufficiently close to be used for estimation. For large crowns, the number of branches can be determined by a sampling procedure, several methods have been described in literature (Schopfer 1981, Valentine et al. 1984).

In the present study the crowns were relatively small, and branches easily could be counted, a sampling procedure to determine the total number would have been more time consuming than complete enumeration. Sample branches had to be selected for deriving the regression estimates. From a sufficiently large sample, regression estimates are derived for single branches with branch weight as dependent and branch base diameter and branch length as independent variables.

The branches to be weighed were selected by PPS sampling, the probability of selection was determined by the dimension of the branch. Previous observation suggested that branch dimension (diameter and length) for this population was closely related to stem diameter (i.e. the location of the branch in the crown). This assumption may not be true for other population.

The probability of selection, therefore, was determined proportional to the stem diameter, the sampling process was conducted in two stages. In the first stage the location along the stem was determined randomly with a probability proportional to stem diameter, the next closest node to this point was selected, and from this node a branch was randomly chosen for measurement. The procedure can be described as two stage sampling with PPS selection on the first stage and an unequal number of secondary units (branches) per primary (node).

The regression model selected was of the form

$$W = b_0 + b_1 D^2 L$$

where W = branch weight
 D = branch base diameter
 L = branch length

A plot of the data showed that one assumption of ordinary least squares, the homogeneity of variances was not fulfilled. Therefore, a weighted least squares method was used.

The regression coefficients for estimating branch weight are listed in Table 2.

Table 2.--Regression coefficients for estimating branch weight ($W = b_0 + b_1 D^2 L$)

Tree	b_0	b_1
1	-.00593	.12396
2	.0555	.11149
3	-.01959	.12967
4	-.00962	.12637
5	.03441	.09492
6	-.02668	.10536
7	-.09962	.13863
8	.06554	.05411
9	.07806	.09999
10	.01860	.12178
11	.03586	.08597
12	.02767	.11980
13	.08619	.06848
14	.04392	.10792
15	.09882	.09027
16	.10094	.10064

The branch base diameters were determined only for a small sample, therefore, these distributions were smoothed before they were converted to a entire tree basis.

The total fresh weight of the crowns were estimated by applying these single tree equations to the tree data. The results (Table 3) show that there is a negative bias of about 11% for the 16 trees, probably in part due to the fact that small branches not in a node have not been counted.

Additional studies are being made to show the source of this bias and to find a sampling procedure more accurate. The dry weight subsequently was estimated by taking a subsample of 5 branches from each tree that were dried at 103° C for 72 hours, simple ratio estimates were used to convert to ATRO weight.

Combining the estimation of bole and crown weight (Table 4), the total tree weight (above ground) could be estimated with a negative bias of about 5%.

Further studies will be conducted to develop more accurate and efficient methods to estimate crown weight as sampling crown weight can result in significant reductions of biomass inventory cost.

Table 3.--Accuracy of estimating crown weight

Tree	Crown Weight (kg)	
	True	Estimated
1	34.7	33.0
2	20.7	16.0
3	14.3	13.5
4	36.9	34.7
5	20.8	19.0
6	36.7	24.1
7	61.8	56.7
8	16.8	11.2
9	41.7	36.4
10	25.8	26.8
11	13.8	8.9
12	25.1	27.5
13	21.1	18.5
14	40.8	34.6
15	16.6	17.1
16	36.1	38.5
Total	463.7	.5

Table 4.--Accuracy of estimating total weight

Tree	Total Weight (kg)	
	True	Estimated
1	162.6	146.8
2	93.8	81.6
3	77.8	73.2
4	151.2	148.1
5	93.1	86.8
6	165.6	149.5
7	216.0	216.2
8	84.0	77.9
9	162.1	165.4
10	89.7	91.5
11	65.3	57.9
12	110.3	109.7
13	105.6	96.9
14	173.6	161.3
15	93.0	95.3
16	172.3	164.4
Total	2016.0	1922.5

Conclusions

Total above ground biomass may be estimated by sampling within trees. The sampling procedure suggested for estimating stem weight is a selection proportional to volume (PPS), the stem parts with higher volumes are sampled more heavily. The results of this estimation procedure were satisfactory for the 16 trees studied, suggesting that direct total weight measurements of the stem could be substituted by sampling. No separation of wood and bark was made in this study.

For estimating total crown biomass branches were sampled, with a selection probability approximately proportional to the size of the branch (for the given population stem diameter instead of branch diameter or length could be used for determining selection probabilities). The total number of branches were counted, total crown biomass was calculated based on this weighed sample, the accuracy was quite satisfactory, suggesting that total tree biomass can be accurately predicted by sampling instead of weighing all the components.

Literature Cited

- Cunia, T. 1979. On sampling trees for biomass table construction: some statistical comments. In: Forest resource inventories Vol. 2, p. 643-664. Colorado State University, 1030 p.
- Schopfer, W. 1961. Beitrage zur Erfassung des Assimilationsapparates der Fichte. Schriftenreihe Ba-Wu FVA Abt. Biometrie 10.
- Valentine, H.T.; Trittonm, L.M.; Furnival, G.M. 1984. Subsampling trees for biomass, volume, a mineral content. For. Sci. 30: 673-687.
- Weber, M. 1983. Inventurverfahren zur Biomassen ermittlung von Waldbestanden. Thesis Abt. Forstl. Biometrie Univ. Freiburg.

SIMPLE BIOMASS REGRESSION EQUATIONS FOR

SUBTROPICAL DRY FOREST SPECIES

Joseph D. Kasile

Associate Professor of Natural Resources
The Ohio State University
Columbus, OH 43210

Simple linear regression equations were developed for the sixteen major forest species of the dry forests of the Dominican Republic. Diameter at 0.5 meters above ground is the easiest to use independent variable and total green biomass and usable green biomass are the dependent variables. Usable biomass is that portion of the tree above the stump and greater than 2.5 centimeters, which are the common usage limits for firewood and charcoal production. Correlation coefficients squared of about 0.95 were achieved for most species.

Subtropical Dry Forests

The subtropical dry forests of the Dominican Republic (DR) have long served as the most important fuelwood source for the rural poor of the nation and are now being considered by the Dominican government as a supplemental energy source for the country. Comprising approximately 5,905 square kilometers or 12 percent of the nation's land area (Jennings and Ferrerías 1979), the subtropical dry forests are a major resource to the Dominican Republic. To establish sound management practices and policies for that nation's dry forests, it is necessary to adapt a forest inventory methodology that will assess the total above ground forest biomass and the biomass usable for fuelwood.

The vast majority of the biomass regression equations reported to date are for predicting green and dry weights of tree components (leaves, branches, stems and/or roots) for single stem broadleaf or conifer species in the temperate regions. Very few biomass studies have been conducted in the subtropical dry forests and no regression equations have been developed that predict the biomass weight of the species found within these dry forest ecosystems where the scarcity of fuelwood is most acute and the utilization of the tree components above ground is most common.

Sample Data

The field data for the biomass equations were collected from June - September, 1984. Sample trees were selected randomly throughout the 850 hectare experimental dry forest in Mao, DR. The twenty species found in the northwestern subtropical dry forest of the Dominican Republic

that are considered important for firewood or charcoal making and/or are abundant in the forest were sampled. The common, scientific and family names of the twenty species are found in Table 1.

Table 1

Common, scientific and family names of species included in the biomass study of the dry forest of Mao, Dominican Republic

<u>Common Name</u>	<u>Scientific Name</u>	<u>Family Name</u>
Almácigo	<u>Bursera simaruba</u>	Burseraceae
Aroma	<u>Acacia macracantha</u>	Mimosaceae
Baitoa	<u>Phyllostylon brasiliensis</u>	Ulmaceae
Bayahonda	<u>Acacia tortuosa</u>	Mimosaceae
Brucón	<u>Cassia emarginata</u>	Caesalpinaceae
Cafetán	<u>Palicourea alpina</u>	Rubiaceae
Cambrón	<u>Prosopis juliflora</u>	Mimosaceae
Candelón	<u>Acacia scleroxyla</u>	Mimosaceae
Ciguamo	<u>Krugiodendron ferreum</u>	Rhamnaceae
Cinazo	<u>Pithecellobium circinale</u>	Mimosaceae
Frijol	<u>Capparis</u> spp.	Capparidaceae
Guaconejo	<u>Amirys</u> spp.	Rutaceae
Guatapanal	<u>Caesalpinia coriaria</u>	Caesalpinaceae
Guayacán	<u>Guaiacum officinale</u>	Zygophyllaceae
Mostazo	<u>Capparis flexuosa</u>	Capparidaceae
Quina	<u>Exostema caribaeum</u>	Rubiaceae
Sangretoro	<u>Maytenus buxifolia</u>	Celastraceae
Tabacuelo	<u>Pictetia spinifolia</u>	Fabaceae
Trejo	<u>Rhynchosia uncinata</u>	Fabaceae
Uvero	<u>Coccoloba leoganensis</u>	Polygonaceae

All sample trees had the following measurements taken:

1. Diameter at 0.5 meters above ground level recorded to the nearest 0.1 centimeter (DKH). This diameter is referred to as "diameter at knee height."
2. Diameter at 1.3 meters from the root collar recorded to the nearest 0.1 centimeter (DBH).
3. Crown diameter recorded to the nearest 0.1 meter (CD). Before felling the tree, this variable was measured with a metric tape. The average of two measurements, one in the north-south and the other in the east-west direction, was recorded on the tally sheet. In all directions the outer edge of the crown was considered to be the last green leaf.
4. Total height recorded to the nearest 0.1 meter (TH). After felling the tree with an axe, the total height was measured. To account for variable stump heights, the total height measurement was made from the point on the stem previ-

ously marked at 0.5 meters to the last green leaf on the crown and then adding 0.5 meters.

5. Green weight of the total live biomass above ground recorded to the nearest pound (GWT).

6. Green weight of the usable biomass recorded to the nearest pound (GWU). This is all the aerial parts of the tree having a diameter greater than or equal to 2.5 centimeters. This minimum diameter is generally considered by the Dominican forestry personnel to be the smallest diameter used by the local people for firewood or for making charcoal.

A circular spring scale was used to measure the green weight of the trees cut in the field. The scale was hung from a portable metal tripod. A collapsible 1.5 meter and 1.5 meter square platform was then hung with nylon ropes from a hook on the bottom of the scale on which the biomass was supported.

Distinctions were made between:

- all living aerial portions of the tree (GWT) and
- all living woody parts of the tree with a diameter greater than 2.5 centimeters (GWU).

Results And Discussion

For each specie, the average dimensions measured for the trees in the field are shown in Table 2.

The first two columns of Table 2 show the number of stems that were measured at 0.5 and 1.3 meters above ground level, respectively. The comparison of these two columns gives an indication of those species whose stems may be more likely to fork between 0.5 and 1.3 meters.

For the twenty species combined, there are approximately 1300 stems per hectare having a DBH greater than or equal to 4.0 centimeters. Figure 1 shows that nearly half of these stems have a DKH that falls in the 4-6 centimeter range. The frequency of small stemmed trees illustrates the typical configuration of the dry land forests of this region with many small stemmed trees per hectare. The mean DKH of all stems measured in the sample data was 7.8 centimeters. The largest DKH measured was 34.0 centimeters for a single stemmed *Bursera simaruba*.

The mean height of all the stems measured in the sample data was 6.0 meters. Figure 2 illustrates that more than sixty percent of the stems measured had total heights between 5 and 7 meters indicating the relatively low canopy of the dry land forests. The tallest tree measured was the same *Bursera simaruba* as mentioned above with a total height of 10.5 meters.

Table 2

Summary statistics for 20 species measured in the dry forest Mao, Dominican Republic

Specie	Number of stems measured at:		DKH range		DBH range		Hgt range		Percentage of Usable Biomass (%)
	0.5 m.	1.3 m.	cm.	cm.	cm.	cm.	m.	m.	
<i>Bursera simaruba</i>	20	20	5.5-34.0	15.5	4.5-32.0	14.6	3.8-10.5	7.0	87
<i>Acacia macracantha</i>	3	4	5.4-14.7	9.0	4.9-10.9	7.2	5.5-7.2	6.6	75
<i>Phyllostylon brasiliensis</i>	104	115	4.3-18.5	7.1	3.4-17.4	5.9	4.1-9.6	6.1	64
<i>Acacia tortuosa</i>	1	1	6.8	6.8	5.3	5.3	6.3	6.3	78
<i>Cassia emarginata</i>	11	11	4.1-11.6	6.3	4.1-10.5	5.9	5.5-6.6	6.0	60
<i>Palicourea alpina</i>	15	18	4.8-12.6	6.2	4.0-10.3	5.1	4.4-6.6	5.3	57
<i>Prosopis juliflora</i>	12	12	4.5-15.0	6.8	4.2-10.9	5.9	4.2-8.0	6.1	75
<i>Acacia sclerosia</i>	25	33	4.7-28.8	9.5	4.0-20.8	7.7	4.7-9.8	6.4	71
<i>Krugiodendron ferreum</i>	2	2	5.0-5.7	5.4	4.6-4.7	4.7	4.7-4.8	4.8	59
<i>Pithecellobium circinale</i>	10	10	4.1-8.1	5.6	4.1-6.8	5.2	5.7-6.5	6.1	63
<i>Capparis</i> spp.	2	4	12.0-13.0	12.5	7.6-10.5	9.0	4.9-5.6	5.3	73
<i>Amirys</i> spp.	7	9	4.6-9.9	6.8	4.1-8.4	5.8	4.3-6.4	5.0	55
<i>Caesalpinia coriaria</i>	3	4	9.2-12.5	10.8	8.3-10.3	9.0	4.7-6.4	5.5	55
<i>Gualicum officinale</i>	13	15	4.8-8.3	6.3	3.5-7.2	4.9	3.2-5.5	3.9	69
<i>Capparis flexuosa</i>	5	6	4.3-6.9	5.3	3.4-5.4	4.5	3.8-6.1	5.0	48
<i>Exostema caribaeum</i>	18	20	4.7-11.2	6.7	3.0-8.1	5.2	5.6-8.0	6.7	74
<i>Maytenus buxifolia</i>	2	3	7.0	7.0	4.2-5.5	4.7	3.4-4.6	4.0	57
<i>Pictetia spinifolia</i>	3	3	5.5-7.0	6.0	4.3-5.7	5.0	3.6-5.5	4.6	79
<i>Rhynchosia uncinata</i>	1	1	4.6	4.6	4.2	4.2	3.7	3.7	62
<i>Coccoloba joganensis</i>	1	1	5.7	5.7	5.5	5.5	5.3	5.3	45
TOTAL	258	292	4.1-34.0	7.8	3.0-32.0	6.6	3.2-10.5	6.0	71

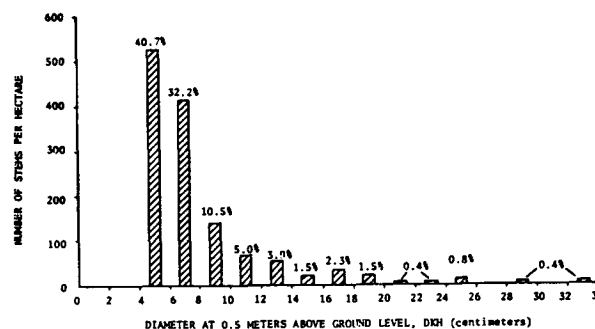


Figure 1. Diameter distribution for stems with a DBH greater than or equal to 4.0 centimeters in ten 200 m² sample plots of the Mao subtropical dry forest.

The percentage of the total green weight biomass that is usable for firewood or making charcoal is also shown in Table 2. The ratio of the usable to total green weight biomass shown in Table 2, estimates what percentage of the total above ground portion of the stem for a particular species is considered usable for firewood or

making charcoal. High percent values such as those of *Bursera simaruba*, *Acacia macracantha* and *Prosopis juliflora* indicate that if used for fuelwood, the major part of the above ground portion of the stem would normally be utilized. Low percent values, on the other hand like those of *Coccoloba leoganensis*, *Capparis flexuosa* and *Caesalpinia coriaria* demonstrate that a large part of each of these species total above ground biomass are of the crown portions having a diameter less than 2.5 centimeters. The above ground portions having a diameter less than 2.5 centimeters are considered "unusable" biomass for firewood or charcoal making.

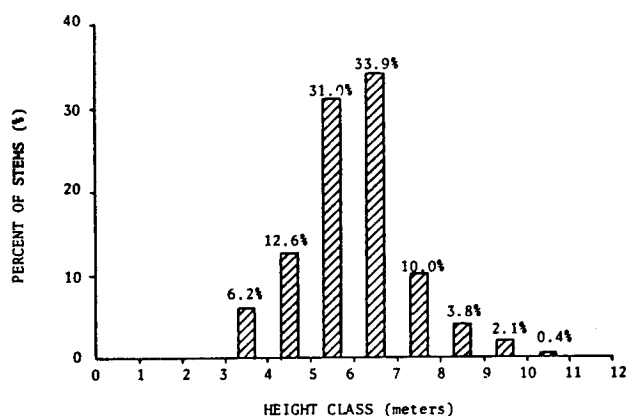


Figure 2. Height distribution for stems with a DBH greater than or equal to 4.0 centimeters in the ten 200 m² sample plots of the Mao subtropical dry forest.

Biomass Regression Equations

Tables 3 and 4 present the regression coefficients and sample statistics of the prediction equations to estimate the green weight of the usable biomass or the portions having a diameter greater than or equal to 2.5 centimeters for sixteen individual species and three combined species. During the development of these equations a double logarithmic transformation was used to ensure that the variances were homogeneous.

Table 3 presents the simple linear regression equations, with the independent variables that yield the highest r². Seven of the ten independent variables considered, give the highest r² value in all seventeen biomass equations. Each equation in Table 3 is of the form:

$$\ln(\text{weight}) = b_0 + b_1 \ln(X_i)$$

where:

weight=green weight in pounds of the usable biomass.

X₁:X₁=basal area of stem at 0.5 meters above ground level, (DKHBA).

X₂=basal area of stem at 0.5 meters above ground level times total stem height, (DKHBA * TH).

Table 3

Simple linear regression coefficients and statistics for predicting green weight of the usable biomass of subtropical dry forest species in Mao, Dominican Republic with the regression model:

$$\ln(\text{usable weight}) = b_0 + b_1 + \ln(X_i)^{1/}$$

Species ^{2/}	Coefficients		X _i	Statistics			
	b ₀	b ₁		n	Sy.x ^{3/}	r ²	DKH Range (cm)
<i>Bursera simaruba</i> Almacigo	7.191616	1.110368	X ₂	36	0.146939	.9898	4.0-34.0
<i>Acacia macracantha</i> Aroma	7.633640	1.043212	X ₃	23	0.183606	.9831	4.0-30.0
<i>Phyllostylon brasiliensis</i> Baitoa	7.096829	1.038230	X ₂	56	0.144541	.9849	4.0-29.0
<i>Cassia emarginata</i> Brucon	10.008309	1.208762	X ₁	26	0.178667	.9574	4.0-16.0
<i>Policurea alpina</i> Cafetan	9.592580	1.125651	X ₁	18	0.277910	.8287	4.0-12.0
<i>Prosopis juliflora</i> Cambon	7.518781	0.999669	X ₆	22	0.168407	.9827	4.0-30.0
<i>Acacia sclerosyla</i> Candelon	7.762502	1.112649	X ₆	46	0.173730	.9854	4.0-29.0
<i>Pithecellobium circinale</i> Cinazo	8.804170	1.315341	X ₆	23	0.211858	.8732	4.0-8.0
<i>Capparis</i> spp. Frijol	7.334954	1.053208	X ₂	21	0.116820	.9880	4.0-19.0
<i>Amirys</i> spp. Guaconejo	7.661575	1.111198	X ₂	21	0.111727	.9759	4.0-11.0
<i>Caesalpinia coriaria</i> Guatapanal	7.882536	1.079668	X ₅	24	0.184299	.9822	4.0-24.0
<i>Guaiacum officinale</i> Guayacan	11.386864	1.464382	X ₁	24	0.201363	.9156	4.0-10.0
<i>Exostema caribaeum</i> Quina	4.797154	0.820955	X ₆	29	0.166601	.9528	4.0-14.0
<i>Maytenus buxifolia</i> Sangreoro	7.850758	1.120838	X ₂	22	0.180072	.9516	4.0-14.0
<i>Pictetia spinifolia</i> Tabacuelo	7.390734	1.088975	X ₂	13	0.177935	.8886	4.0-7.0
<i>Coccoloba leoganensis</i> Uvero	5.701871	0.974551	X ₆	26	0.189499	.9657	4.0-18.0
Others	6.480793	1.157791	X ₇	11	0.285666	.8381	4.0-8.8

^{1/} X_i = independent variable where:

X₁ = DKHBA X₃ = DKHBA * CD X₅ = DBHBA * CD X₇ = DBHBA * TH²
X₂ = DKHBA * TH X₄ = DBHBA * TH X₆ = DKHBA * TH²

^{2/} Common name of species as referred to in northwestern Dominican Republic
^{3/} Standard error of the estimate in natural log (ln) form.

X₃=basal area of stem at 0.5 meters above ground level times crown diameter, (DKHBA * CD).

X₄=basal area of stem at 1.3 meters above ground level times total stem height, (DBHBA * TH).

X₅=basal area of stem at 1.3 meters above ground level times crown diameter, (DBHBA * CD).

X₆=basal area of stem at 0.5 meters above ground level times total stem height squared, (DKHBA * TH²).

X₇=basal area of stem at 1.3 meters above ground level times total stem height squared, (DBHBA * TH²).

The umbrella-like crown canopy and multi-stemmed nature of many species in the natural dry forests make it difficult, time consuming and expensive to accurately measure the total height or crown diameter of a standing tree. Therefore, the regression equations shown in Table 4 for each specie, predict the green weight of the usable biomass with the independent variable being the basal area of the stem at 0.5 meters

Table 4

Simple linear regression coefficients and statistics for predicting green weight of the usable biomass of subtropical dry forest species in Mao, Dominican Republic with the regression model:

$$\ln(\text{usable weight}) = b_0 + b_1 + \ln(\text{DKHBA})^{1/}$$

Species ²	Coefficients		n	S _{y.x} ³	Statistics r ²	DKH Range (cm)
	b ₀	b ₁				
<i>Bursera simaruba</i> Almacigo	10.304117	1.354212	36	0.201718	.9807	4.0-34.0
<i>Acacia macracantha</i> Aramo	10.926375	1.312754	23	0.286349	.9590	4.0-30.0
<i>Phyllostylon brasiliensis</i> Baitoa	10.119439	1.243693	56	0.175208	.9778	4.0-29.0
<i>Cassia emarginata</i> Brucón	10.008809	1.208762	26	0.178667	.9574	4.0-16.0
<i>Pellicurea alpina</i> Cafetan	9.592580	1.125651	18	0.277910	.8287	4.0-12.0
<i>Prosopis juliflora</i> Cambrón	9.803672	1.137246	22	0.288090	.9493	4.0-30.0
<i>Acacia sclerosyla</i> Candelón	11.106018	1.409781	46	0.270065	.9646	4.0-29.0
<i>Pithecellobium circinale</i> Cinazo	11.074430	1.320648	23	0.270181	.7938	4.0-8.0
<i>Capparis</i> spp. Frijol	10.047913	1.272722	21	0.149780	.9803	4.0-19.0
<i>Amirys</i> spp. Guaconejo	10.624340	1.323345	21	0.127177	.9688	4.0-11.0
<i>Casahuate coriaria</i> Guatapán	11.733678	1.575985	24	0.276606	.9600	4.0-24.0
<i>Gusticium officinale</i> Guayacán	11.384864	1.464382	24	0.201363	.9156	4.0-10.0
<i>Exostema caribaeum</i> Quina	9.717862	1.171786	29	0.285535	.8613	4.0-14.0
<i>Maytenus buxifolia</i> Sangre toro	10.699403	1.357645	22	0.237306	.9160	4.0-14.0
<i>Picetia spinifolia</i> Tabacuelo	11.290882	1.495389	13	0.212988	.8403	4.0-7.0
<i>Coccoloba leoganensis</i> Uvero	10.226620	1.189277	26	0.292904	.9181	4.0-18.0
Others	11.351772	1.393268	11	0.433486	.6271	4.0-8.8

¹/DKHBA is the basal area at 0.5 meters above ground level expressed in square meters (m²).
²/Common name of species as referred to in northwestern Dominican Republic
³/Standard error of estimate in natural log (ln) form.

above ground level, DKHBA. The basal area at knee height was chosen as the independent variable because it showed the highest correlation to stem biomass for most species. Also DKHBA is the fastest and easiest variable to measure of all the variables considered. The equations in Table 3 have smaller standard error of the estimates and higher r² as compared to those of Table 4 which indicate that Table 3 regression equations will give a more precise estimate of the usable and total biomass for stems within the specified diameter ranges of each species. However, in using equations of Table 3 another variable (height, crown diameter or diameter at breast height) must be measured in the field for each stem. This increases the amount of necessary field work and the overall cost of the sampling procedures. Also, in comparing Table 3 with Table 4 in most cases, the increase in the r² from using a variety of independent variables as required by Table 3 results in little improvement of the r² over using DKHBA as a uniform independent variable. For inventorying stands with mixed species, probably the set of equations having the same independent variable, i.e.,

DKHBA, would be best. In cases where just one or two species are of interest, however, it might be more feasible to use the equations yielding the most precise estimates.

Summary

Regression equations with simple to measure independent variables were developed for the 19 dry land forest species that will serve as the basic parameters for the inventory process.

These regression equations utilize a single parameter of measurement; the easiest of these to field measure is diameter at 0.5 meters (DKHBA) as the prediction variable of biomass for each species. DKHBA is a new concept of measurement that allows fewer measurements per tree than the standard diameter at 1.3 meters used for single-stemmed trees of the temperate zone. DKHBA is well above root collar swell for these small diameter trees and below most of the thorny branches that make measurement difficult. Almost all the squared correlation coefficients (r²) were .90 or more.

Using the biomass regression equations shown in Tables 3 and 4 on inventory data will provide accurate biomass estimates of the material commonly used for firewood and making charcoal in the native dry forests of the Dominican Republic.

Inventories of the dry land forests can now be conducted in an efficient, accurate manner to determine biomass for each area. These areas can then be regulated to provide the continuing, renewable fuelwood energy needs of the Dominican Republic.

Literature Cited

- Jennings, P. and B. A. Ferreira. 1979. Recursos energéticos de bosques secos en la República Dominicana. Instituto Superior de Agricultura, Santiago, República Dominicana. 119 p.
- Mercedes, J., M. Powel and J. A. Núñez. 1984. Estudio de la densidad y la variabilidad de las especies del bosque seco en un muestreo de transectos en Mao, República Dominicana. (In press).
- Murphy, P. G. and A. E. Lugo. 1985. Structure and biomass of a subtropical dry forest in Puerto Rico. Biotropica, accepted for publication.

RESEARCH PAPERS

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

Moderator: Eric H. Wharton

EVALUATING ERRORS OF TREE BIOMASS REGRESSIONS BY SIMULATION

Tiberius Cunia

Professor of Statistics and Operations Research
SUNY College of Environmental Science and Forestry
Syracuse, NY 13210

To construct tree biomass regression functions for use in forest biomass inventory, one would normally (i) select a sample of trees by some random procedure, (ii) measure the biomass of the sample trees, and (iii) calculate a biomass regression function by some statistical procedure. The sampling method is generally complex and its structure is seldom taken into account by the estimation procedure. To see the effect of sampling method and estimation procedure on the error of the inferences made when the tree biomass regressions are used, simulation techniques were applied to a population of 22,753 forest trees distributed over 927 one-fifth acre plots. Samples of trees were repeatedly selected by different methods and, for each sample, biomass regressions were calculated by various statistical procedures. Applied to the parent population, the biomass regressions generated estimates of the average biomass per unit area and their error. From the analysis of the differences between true and estimated average biomass, conclusions were drawn about the bias, precision, and estimated precision of combinations of sampling method and estimation procedure.

Introduction

The sampling designs used in forest biomass inventory consist generally of two phases. In the first phase, a relatively large sample of trees (usually in clusters defined as trees growing on plots of fixed area or Bitterlich sample points) is selected by some statistical sampling procedure. These trees are measured for species, diameter, height, etc. but not measured for biomass. In the second phase, a relatively small sample of trees is randomly selected and its trees are measured for biomass in addition to species, diameter, height, etc. A tree biomass regression function (on species, diameter, etc.) is calculated from the second phase sample trees and this regression function applied to the trees of the first phase yields estimates of the average biomass per unit area or total biomass in the entire forest area.

Because of the structure of the sampling design, the error of the biomass estimates has two main components; one component due to the error of the first phase and a second component due to the error of the tree biomass regression function of the second phase. This last component is usually ignored when the error of the biomass estimates is calculated; it is difficult to assess the error of biomass regression in meaningful terms and although

the methodology to combine the two components exists, it is largely unknown to the forest mensurationists.

The objectives of the present study are those of evaluating the error of biomass regression functions as applied to forest inventory data for the purpose of estimating the average biomass per unit area. For this, the error of the biomass regression should be expressed in a suitable form; it must be combined with the error of the first phase sample data when the error of the average biomass per unit area is being estimated. In our study we shall assume that, to combine the errors from the two phases we shall assume that the approach suggested by Cunia (1965, 1986) is being used. This approach requires that (i) the biomass regression functions be of the linear form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

where y = tree biomass, $x_1 = 1$, and $x_2, x_3, \dots, x_m$ are the independent variables defined in terms of diameter, height, species, etc, and (ii) valid estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix $[\sigma_{bb}]$ of $[b]$ exist and are known. Consequently, the error expression of the biomass regression function used in our study is $[S_{bb}]$.

Because the error of the regression is a function of (i) the sampling design by which the sample trees are selected, and (ii) the estimation procedure by which the regression is calculated, we shall consider both these factors when evaluating this error. We shall not consider, however, (i) the error of measurement of the biomass of the sample trees when determined by subsampling; we shall assume that the trees are measured without error, (ii) the error of the first phase sample; we shall apply the biomass regression to the entire tree population, and (iii) the error made when the trees selected from one are applied to another tree population.

To attain our objectives, we have used computer simulation techniques consisting of the following main steps. We have started with the construction of a real world type population of some 23,000 trees distributed in clusters (sample plots of fixed area) and measured, among other things, for diameter, height and biomass. This population was stored in the computer. Using a variety of sampling methods, sample trees were then selected by computer and their diameter, height, and biomass recorded. There were 100 repetitions of this entire simulated sampling process resulting in 100 samples of trees, hopefully independent, for any given sampling method and sample size. In the next step, biomass regression functions were calculated by a variety of estimation procedures, one set of regressions for each sample of trees generated by the previous step. Each tree biomass regression was then applied to the tree population to estimate the known value of the average biomass per unit area. Finally, the hundreds of thousands of estimated values were analyzed for possible bias, error and efficiency of each combination of sampling method and estimation procedure.

Construction of the Tree Population

The basic data used to construct our tree population consists of field measurements of diameter at breast height d , merchantable height h_m , and species, among other things, performed on 22,753 trees contained in 927 one-fifth acre plots selected by some random procedure from the New York State forest lands. Only the merchantable trees were included in the population (trees of diameter not less than 5 inches) and 260 of the 927 sample plots were empty, that is, did not contain trees of merchantable size. The total height h and the total above ground biomass y (green weight) was not measured in the field but was generated for each population tree by a Monte Carlo procedure described in detail in a series of papers by Cunia and Michelakackis (1983b, 1984a, b) and Cunia, Michelakackis and Lee (1984). Because the procedure by which h and y was generated is an important factor to consider when interpreting the results of the analysis of the simulated sample data, the method to generate h and y is summarized below as follows.

The total height of each tree was generated first by a formula of the form $h = \hat{h} + q$ where (i) h is the simulated total tree height, (ii) $\hat{h}$ is the regression estimate of the conditional mean of height for given species, diameter d , merchantable height h_m , and geographical area, and (iii) q is a random variable expressing the difference between actual tree height and average height as estimated by the regression function. The regression function of total height on diameter and merchantable height was estimated from actual sample data of some 1,600 trees of twenty species selected from three states, New York, Michigan and West Virginia. The least squares method was used and the regression function was assumed to be of the linear form $\hat{h} = b_1 + b_2 d + b_3 h_m$. Separate regression functions by species and state were calculated and most of the time either d or h_m was sufficient to adequately define the regression relationship. This means that for most regressions, either b_2 or b_3 was made equal to zero. The regression coefficients b_1 , b_2 , and b_3 were stored in the computer and applied to our tree population data to calculate the values $\hat{h}$.

The random variable q was generated by a Monte Carlo technique according to a probability distribution estimated from the same sample trees. It was assumed that the shape but not the scale of the probability distribution of q was the same for all species. If (i) e is a standardized random variable with mean zero and variance one, and (ii) S_{hh} is the conditional variance of total tree height for given diameter, merchantable height, species and state, the random variable q was defined as being equal to $e\sqrt{S_{hh}}$. The variance S_{hh} was estimated separately for each sample regression but the probability distribution of $e = q/\sqrt{S_{hh}}$ was estimated from the pooled data of all regressions. As the sample distribution was found to be irregular in shape, a two-stage graphical procedure was used to smooth it out.

Using the total height h as generated above, the biomass of each population tree was generated

by a similar procedure. The formula used was of the form $y = \hat{y} + q$ where (i) y is the value of the tree biomass obtained by this simulation process, (ii) $\hat{y}$ is the regression estimate of the conditional mean of biomass for given species, diameter d , total height h , geographical region and cluster (plot) and (iii) q is a random variable expressing the difference between actual tree biomass and the conditional mean biomass as estimated by the regression function. The regression function of y on d and h was estimated from actual sample tree data from Finland, New York, Michigan, Ohio, and West Virginia. The weighted least squares method was applied to a regression function of the linear form $\hat{y} = b_1 + b_2 d^2 h$ under the assumption that, within each species and state, the conditional variance of y is proportional to $(d^2 h)^2$. The regression functions were calculated separately by species and state and the effect of site was estimated by regression with dummy variables techniques. To take the cluster effect into account, we have estimated the probability distribution of pairs of regression coefficients b_1 and b_2 calculated within each individual cluster (plot) from their own tree data.

To estimate the probability distribution of q , we have assumed that $q = e(d^2 h)\sqrt{S_{uu|y}}$, where (i) e is a standardized random variable with mean zero and variance one, (ii) $S_{uu|y}$ is the estimate of the conditional variance of the transformed variable $u = y/d^2 h$ for given diameter and height, and (iii) the conditional standard deviation of y given diameter d and height h is estimated by $d^2 h\sqrt{S_{uu|y}}$. The values of $S_{uu|y}$ were calculated separately by species and state and the probability distribution of $e = q/(d^2 h)\sqrt{S_{uu|y}}$ was estimated from the pooled data of all regressions. A graphical two-stage procedure was also used to smooth out the irregular shape of the sample probability distribution of e .

To generate the biomass of a given tree, the Monte Carlo technique proceeded as follows. For the trees of a given plot a set of regression coefficients b_1 and b_2 was generated for each species according to the bivariate probability distribution of the pairs of plot values b_1 and b_2 . Using the tree diameter, height, species and the geographical region of the plot, the regression value $\hat{y}$ was then calculated by the corresponding regression function. Finally, a value $q = e(d^2 h)\sqrt{S_{uu|y}}$ was added to $\hat{y}$, where (i) d and h are the tree diameter and height, (ii) $\sqrt{S_{uu|y}}$ is the stored value of the conditional standard deviation of $u = y/d^2 h$ for the given species and region, and (iii) e is a random variable with mean zero and variance one generated by the computer according to the probability distribution of e also stored in the computer.

Sampling Procedures

Although easily done by computer, it is relatively difficult in real life to select trees individually by simple random sampling. The usual procedure is to select the sample trees in clusters. Ordinarily defined as groups of trees contained in plots of fixed area, or trees counted by relascope from randomly selected points in the forest, the clusters may be selected by simple or stratified

random sampling, in one or two-phases and all the trees, or a randomly selected part of the trees of the sample clusters are measured for biomass, diameter, height and other attributes of interest. In our simulation study we have sampled trees by three main sampling procedures, denoted here, for convenience, as (i) two-stage random sampling, (ii) two-stage stratified sampling, and (iii) two-phase, two-stage random sampling. Within each sampling method we have defined several variants by changing the tree subsampling procedure of the second stage. Here is a more detailed description of our three sampling methods.

(1) Two-stage random sampling. In the first stage, m plots (clusters) are selected from the 667 non-empty plots of our population by simple random sampling without replacement. In the second stage, sample trees from the sample plots of the first stage are selected by one of seven subsampling procedures. In the first procedure, a fixed percentage p of trees is selected by simple random sampling without replacement. If a fixed number r of trees is selected, subsampling methods 2 and 3 are obtained when the selection is made without or with replacement respectively. The last four subsampling methods are obtained when the tree selection is made with replacement and with probability proportional to a measure of tree size; more specifically, proportional to tree height h , tree diameter d , tree basal area (d^2) and approximate tree volume (d^2h) respectively for methods 4, 5, 6, and 7.

The size of the sample is controlled by the number m of sample plots and the percentage p , or the fixed number r of trees subsampled within the plot. We have used the values $m = 1, 2, 5, 10, 15, 20, 30$, and 50, the values $p = .05, .10, .15, .30, .40, .60$ and 1.00 and the values $r = 1, 2, 5, 10, 15, 20$, and 30. In a second simulation study we have also used $p = .10, .20, \dots, 1.00$. Because we did not want samples that were too small or too large, we have used only those combinations of m with p or r for which the expected sample size fell between 10 and 500.

Note that the only sampling method resulting in selection of trees with equal probability from the entire population of trees is subsampling method 1. Selecting a fixed number of trees from each randomly selected plot (as in subsampling methods 2 and 3) results in a larger probability of selection for the big trees. It is known that the larger trees require more living space than the smaller trees and, thus, the big trees have the tendency to be contained in plots with a small, rather than a large number of trees. For the same reason, selecting fixed number of trees with probability proportional to h , d , d^2 , and d^2h (subsampling methods 4, 5, 6, and 7) results in tree selection with probability higher than h , d , d^2 , and d^2h respectively.

To reduce the amount of simulation work required, we have selected the sample plots and the trees within the sample plots in a nested fashion. To better describe this procedure, let us consider, as an illustrative example, the first subsampling method. In the first simulation run, the run that

generates the first sample of all sample sizes, 50 plots are selected by simple random sampling without replacement. Five percent of the trees selected from the first plot constitutes the first sample of size $m = 1$, $p = .05$. By adding five percent more trees from the same plot, we obtain the first sample of size $m = 1$, $p = .10$. But if we add instead five percent of the trees selected from the second plot, we obtain the first sample of size $m = 2$, $p = .05$. Continuing this way (in two directions, adding trees and plots), we finish with the selection of the first sample of the last sample size of $m = 50$, $p = 1.00$. Of course, some of these samples will be discarded; those that result in combinations of m and p for which the expected sample size is too small or too large.

Multiplying p by the number of trees in a given sample plot does not generally result in an integer number of sample trees. To decide whether an additional tree corresponding to the fractional part f (from .01 to .99) is to be selected, a random number R generated by the computer is being used. If $R \leq f$, an additional sample tree selected at random from the remaining unselected trees of the plot is added to the sample. Otherwise, no additional sample tree is selected.

When the subsampling of trees is done with replacement, the same tree may be selected in the sample more than once. This may present a problem when the sample plot has very few trees and the fixed number r is very large. For example, if $r = 20$ and the sample plot has one tree, the same tree will appear twenty times in the sample. To improve the sample, we have instructed the computer to generate a new biomass value for every tree already in the sample that happens to be selected again. The procedure to generate the new biomass is identical to that used when the tree population was constructed.

(2) Two-stage stratified sampling. With this method, the m sample clusters are selected by stratified random sampling. The tree population was first divided into three geographical regions (strata) and a two-stage random sample selected from each stratum separately. The same seven subsampling methods of tree selection from sample plots were used here as well. By combining now samples of various sizes from various strata, we obtain two-stage stratified samples of different sizes and allocations.

(3) Two-phase, two-stage random sampling. In the first phase, the trees from a relatively small two-stage random sample are measured for biomass y , diameter d and height h . The tree data are used to estimate the regression function of biomass on diameter and height, say $\hat{y} = r_1(d, h)$. To reduce the computer simulation work, the samples already selected under the two-stage random sampling method of (1) above were used as the first phase sample of the two-phase, two-stage random sampling method. In the second phase, a relatively large, two-stage random sample of trees is selected and every tree is measured for diameter and height but not for biomass. Only the first tree subsampling method is used with $m = 50, 100, 150, 200, 300$, and 400 and $p = 2.93, 5.86, 8.79$, and 11.72.

The percentage p was selected so as to yield, on the average 1, 2, 3, and 4 sample trees per plot respectively. The data from this sample are used to estimate the regression function of tree height h on tree diameter d , say $\hat{h} = r_2(d)$. The two regression functions are then combined to obtain a regression function of biomass on diameter alone, say $\hat{y} = r(d)$.

To obtain a two-phase two-stage random sample, we must pair a sample of the first with a sample of the second phase. The sample trees of the first phase are selected by seven subsampling methods and there is a variety of sample sizes (values m and p or r) used. The sample trees of the second phase are selected by one subsampling method and there are 24 sample sizes, the product of six values m by four values p . For each simulation run, we obtain one two-phase, two-stage random sample for each combination of subsampling method and sample size of the first phase with a sample size of the second phase. As there are 100 simulation runs, there are 100 samples generated by the same two-phase two-stage random sampling procedure with the same sample size.

Estimation Procedures for the Biomass Regression Functions

Consider first the two-stage random sampling method. For each of the 100 samples generated for the same sample size (m and p or r) and tree subsampling procedure, we have calculated thirty-two biomass regression functions, the combinations of four least squares estimation approaches and eight regression equations.

The first estimation approach is that of the ordinary least squares (OLS) method applied to individual tree data. The fact that several sample trees may have been selected from the same cluster is not taken into account. Because the conditional variance of the tree biomass for given diameter or given diameter and height is not homogeneous, it is more reasonable to use the ordinary weighted least squares (OWLS) method. The variance is assumed to be proportional to d^4 for the regression of biomass on diameter alone, or proportional to d^4h^2 for the regression of biomass on diameter and height. With the OWLS approach the cluster effect is also ignored.

The third and fourth estimation approaches are known here as the modified least squares (MLS) and modified weighted least squares (MWLS) methods. They take into account the cluster effect when the regression function is calculated. This is accomplished by applying the weighted least squares method to the cluster (in our case the plot) not the individual tree values. If the tree variables are denoted by $y, x_1, x_2, \dots$ and if Σ means summation over the trees of a given plot, then the plot variables are defined as $\Sigma y, \Sigma x_1, \Sigma x_2, \dots$. In the MLS approach it is assumed that the conditional variance of Σy is proportional to the number of trees in the plot, while in the MWLS approach the conditional variance is assumed to be proportional to Σd^4 if only d is used in the regression or proportional to Σd^4h^2 if both d and h are used.

The basic justification and methodology of the modified least and weighted least squares is given in Cunia (1979, 1981) and Briggs and Cunia (1982).

For each least squares estimation approach we have used eight regression equations, the first three of biomass on diameter alone, the remaining five of biomass on diameter and height. The equations are all linear, and they contain terms of the form d, d^2, h, dh and d^2h . The significance of the addition of some of these terms was sometimes tested statistically.

Consider now the two-stage stratified sampling method. Using the conclusions reached from the analysis of the eight regression equations applied to the seven subsampling procedures of the two-stage random sampling method, we have selected two equations only, the best equation of biomass on diameter alone and the best equation of biomass on diameter and height. As the four least squares estimation approaches were again used, the total number of combinations of approach and equation is eight.

The trees were selected by stratified sampling and the stratification itself may be used when the regression function is estimated. We have defined nine procedures to take the stratification effect into account. These procedures range from one single regression function calculated for all strata with stratification effect completely ignored to three, separately and independently calculated regressions, one for each stratum. The procedures are not described here, and the interested reader can refer to Arabatzis and Cunia (1986) for a more detailed description. The total number of regression functions calculated for each sample is, thus, equal to 72, the product of two regression equations by four least squares estimation approaches and nine ways to take stratification into account.

With the two-phase, two-stage random sampling methods the estimation procedure is more complex. Twenty regression functions were calculated for each first phase sample, the combinations of four least squares estimation approaches (OLS, OWLS, MLS, and MWLS) and five linear regression equations of the form $\hat{y} = r_1(d, h) = a_1x_1 + a_2x_2 + \dots$, where $x_1 = 1$ and the other variables x are defined in terms of diameter d and height h (d, d^2, h, dh and d^2h). Because the conditional variance of the tree height for given diameter is approximately homogeneous, only two least squares estimation approaches (OLS and MLS) were used with the samples of the second phase. They are combined with four linear regression equations of height on diameter of the form $\hat{h} = r_2(d) = c_1x_1 + c_2x_2 + \dots$ where $x_1 = 1$ and the other variables x are of the form d and d^2 . There is a total of eight regression functions for each second phase sample.

A method to combine a regression function $\hat{y} = r_1(d, h)$ of the first phase with a regression function $\hat{h} = r_2(d)$ of the second phase so as to obtain a regression function $\hat{y} = r(d)$ of biomass on diameter alone has been described by Cunia (1982), Cunia and Michelakakis (1983a), and Michelakakis and Cunia (1985, 1986). This method defines the regression function as

$\hat{y} = r_1(d, \hat{h}) = b_1 + b_2 d + b_3 d^2 + b_4 d^3 + b_5 d^4$
 where the regression coefficients $b_1, b_2, \dots, b_5$ are functions of the coefficients a and c of $r_1(d, h)$ and $r_2(d)$ above and the specific independent variables included in the regression being dependent on the specific regressions $r_1(d, h)$ and $r_2(d)$ used. A method is also described to estimate the covariance matrix of $b_1, b_2, \dots, b_5$ when the regression coefficients a and c and their covariance matrices are given.

To calculate the 160 regression functions of the form $\hat{y} = r(d)$ for each two-phase, two-stage random sample (the 160 combinations of twenty regression functions $\hat{y} = r_1(d, h)$ of the first with eight regression functions $\hat{h} = r_2(d)$ of the second phase) would have been prohibitively expensive in terms of computer time and analysis work. Instead, we have selected the four best regressions of the first phase (one for each of OLS, OWLS, MLS, and MWLS approaches) and the two best regressions of the second phase (one for OLS and one for MLS). Pairing the OLS regression of the second with the OLS and OWLS regressions of the first phase and, similarly, the MLS of the second with the MLS and MWLS regressions of the first phase, results in four regressions of the form $\hat{y} = r(d)$ for each two-phase, two-stage random sample.

Application of the Biomass Regression Functions

The sampling is done for the purpose of deriving an estimate z of the mean biomass per acre μ of our tree population. The value of μ is obviously known. But by analyzing the probability behavior of z , we can make inferences about the bias and precision of a given estimation procedure applied to a given sampling method of a given sample size.

There is one estimate z for each estimation procedure and each simulated sample. If the regression function is denoted in matrix notation as

$$\hat{y} = [b]'[x] = b_1 x_1 + b_2 x_2 + \dots$$

where $x_1 = 1$ and, if the error of this regression function is expressed as the covariance matrix $[S_{bb}]$ of $[b]$, it can be shown that

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

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

where the elements of $[\mu_x]$ are the known expected values of the elements of $[x]$ expressed on a "per acre" basis.

For example, if the regression function is

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

then

$$z = b_1 \mu_{x1} + b_2 \mu_{x2} + b_3 \mu_{x3}$$

where

$$\mu_{x1} = \text{mean number of trees/acre}$$

$$\mu_{x2} = \text{mean (sum of } d) / \text{acre}$$

$$\mu_{x3} = \text{mean (sum of } d^2) / \text{acre}$$

The formulae above are easy to prove. For each tree i of the population, the estimate of its biomass y_i is $\hat{y}_i = [b]'[x_i]$, where $[x_i]$ is the vector of the variables x of the tree i . Adding the estimated biomass of all 22753 trees yields an estimate $\hat{Y} = \sum \hat{y}_i$ of the total biomass $Y = \sum y_i$ of our population. Dividing now $\hat{Y}$ by the total forest area of $A = 927/5 = 185.4$ acres, we obtain an estimate $z = \hat{Y}/A$ of the mean biomass per acre $\mu = Y/A$. Symbolically this can all be written as

$$\begin{aligned} z &= \hat{Y}/A = \sum \hat{y}_i / A \\ &= \sum (b_1 x_{1i} + b_2 x_{2i} + \dots) / A \\ &= b_1 (\sum x_{1i} / A) + b_2 (\sum x_{2i} / A) + \dots \\ &= b_1 \mu_{x1} + b_2 \mu_{x2} + \dots \\ &= [b]'[\mu_x] \end{aligned}$$

The formula of the estimator V of the variance of z follows immediately from the fact that (i) μ_x is known without error and (ii) if $[a]$ is a vector of fixed values and $[v]$ is a vector (of the same order) or random variables with covariance matrix estimated by $[S_{vv}]$, then the variance of $[z] = [a]'[v]$ is estimated by

$$S_{zz} = \sum \sum a_i a_j S_{v_i v_j} = [a]'[S_{vv}][a]$$

Analysis Procedure

For each simulated sample and each regression estimation procedure we have calculated the pair of random variables z and V , estimators of μ and variance of z respectively. The pair of confidence intervals $(z \pm t\sqrt{V})$, where $t = 2.0$ for the 95 percent and $t = 2.6$ for the 99 percent confidence level, were also calculated and whether μ fell below, within or above these intervals was recorded. The 100 simulation runs giving rise to the 100 samples may be viewed as 100 repetitions of a random experiment, each experiment generating two random variables z and V , two random intervals $(z \pm t\sqrt{V})$ and two, trinomially distributed random variables that show whether the confidence intervals include the parameter μ or fall below or above it. For all practical purposes, these 100 repetitions may be viewed as being statistically independent, and we may also assume that from one to the next simulation run, the probability distribution of the random variables and intervals remains the same.

To analyze the probability behavior of z , we have calculated, for each set of 100 simulated samples and for each regression model the following statistics, where $\sum$ stands for summation over the sample values from 1 to 100.

$$\bar{z} = \sum z / 100 = \text{average of the 100 sample values } z \\ = \text{estimator of } \mu$$

$$(\bar{z} - \mu) = \text{estimator of the bias of } z$$

$$\bar{V} = \sum V / 100 = \text{average of the 100 sample values } V \\ = \text{estimator of the variance of } z \text{ under the basic assumptions of the regression model used}$$

$S_{zz} = \sum (z - \bar{z})^2 / 99$ = estimator of the variance of z with no assumptions made other than those of the statistical independence of the 100 simulation runs and no change in the probability distribution of z from one to the next run

$t = (\bar{z} - \mu) / \sqrt{S_{zz}} / 100$ = sample value of t that is used to test the null hypothesis that the bias of z is equal to zero

$\bar{V} / S_{zz}$ = ratio of the two estimates of the variance of z ,

and finally, the number of times μ is found to fall below, within or above the 95 and 99 percent confidence limits defined above.

Note that under the assumptions of (i) statistical independence of the various simulation runs and (ii) unchanging probability distribution of z from one to the next run, the statistics $\bar{z}$ and S_{zz} are unbiased estimators of the unknown mean μ and variance σ^2 of the random variable z . As $\bar{z}$ is used as an estimator of the mean biomass per acre μ , the statistic $(\bar{z} - \mu)$ can be used as an estimator of the true bias of z , say $(\mu - \mu)$. The null hypothesis that this bias is equal to zero can be tested by the t -statistic with 99 degrees of freedom. When $(\mu - \mu)$ is small, the statistic S_{zz} can be used as a measure of the efficiency of the sampling method and estimation procedure.

Note also that in real life only the estimator V of the variance of z is available. This estimator is valid, however, only when the basic assumptions of the regression model are strictly satisfied by the tree population we sample and the tree selection procedure we use. As these assumptions are seldom if ever satisfied, we can make inferences about the goodness of V as an estimator of the variance of z by comparing $\bar{V}$, the average of the 100 sample values V with the unbiased estimator S_{zz} . This comparison can be made by means of sample differences $(V - S_{zz})$, the average difference $(\bar{V} - S_{zz})$ or the sample ratio $\bar{V} / S_{zz}$. We have preferred using the ratio.

The analysis of the bias $(\bar{z} - \mu)$ of z , the precision S_{zz} of z , and the validity of the estimator V has been made in two main steps. In the first step we have compared the various regression models within a given sampling method, tree subsampling procedure and sample size. In this way we were able to identify and, thus, eliminate from further consideration, the least squares estimation approach and the regression equation that may yield consistently poor results, that is, estimators with large bias and/or poor precision.

In the second step we have compared the various sampling methods, subsampling procedures and sample sizes. More specifically, we have made inferences about the effect on bias, precision and estimated precision of z of the factors (i) sample size, that is number of clusters and number of sample trees per cluster, (ii) fixed percentage p or number r of trees selected from sample clusters, (iii) subsampling trees with or without replacement, (iv) subsampling trees with equal

or unequal probability and (v) the overall sampling method, the two-stage random or stratified sampling or two-phase, two-stage random sampling.

To do this analysis, we have constructed hundreds of detailed and summary tables and graphs showing how $(\bar{z} - \mu)$, S_{zz} , $\bar{V}$, $\bar{V} / S_{zz}$, etc. vary with the characteristics of the sampling method and estimation procedure.

Main Conclusions Drawn From the Study

The detailed analysis and the conclusions reached as the result of this analysis are contained in a series of papers by Cunia (1985), Cunia and Gillespie (1985), Michelakackis and Cunia (1985, 1986), Gillespie and Cunia (1986) and Arabatzis and Cunia (1986). We shall not repeat them here. However, it may be of interest to give a summary view of these conclusions, some of which came as a big surprise.

When the regression model is suitably selected, the bias of z is generally small even when it is significantly different from zero. One regression function that generates a large bias is $y = \beta d^2 h$; a function that should usually be avoided. The bias does not seem to be affected by (i) the sample size (number of clusters or number of sample trees per cluster), (ii) the selection of a fixed percentage p or a fixed number of trees from a sample cluster, (iii) whether the sampling is done with or without replacement, or (iv) the least squares regression approach or equation (if properly selected). It seems, however, to be somewhat affected by the probability of tree selection when the clusters are subsampled; it seems to increase from small, non-significantly different from zero when the subsampling is done with equal probability (methods 1, 2, and 3), to slightly larger values when the subsampling is done with probability proportional to h and d (methods 4 and 5) and the bias becomes significantly different from zero when the probability is proportional to tree basal area (method 6) or approximate tree volume (method 7). The bias is also affected by stratification when the allocation of sample trees to strata is poor and the stratification effect is not properly taken into account. The bias does not seem to be affected, however, by the two-phase sampling method considered in this study.

As expected, the precision of z as measured by S_{zz} is affected by (i) the overall sample size; the larger the sample, the smaller the error, (ii) the number of trees per cluster for given overall sample size; the smaller the number of trees, the smaller the error, (iii) the least squares estimation approach; the weighted least squares are slightly better than the least squares approaches, and similarly, the ordinary least or weighted least squares are also slightly better than the corresponding modified approaches, and finally, (iv) the regression functions of biomass on d and h are much better than those on diameter alone.

The fact that the least squares is almost as good as the weighted least squares approach came as a surprise. We did not expect the least squares

method to be that robust. Also as a surprise came the fact that no increase in precision is obtained when the probability of selection moves from equal to proportional to a measure of tree size. On the contrary, for the least squares method (OLS and MLS) the precision seems to decrease dramatically as we move from equal to probability proportional to d (or h), to basal area (d^2), and finally to approximate volume (d^2h).

Not completely unexpected, the precision of z does not seem to be affected by (i) sampling with or without replacement, (ii) selection of a fixed number or a fixed percentage of trees from a sample cluster, and (iii) the form of the regression function when this form is not poorly selected (as for example the form $\hat{y} = \beta d^2h$) or when the independent variables are not the same (when both d and h are used we have, as expected, a much better estimator than when only d is used).

The precision of z can also be estimated by V . As expected, the precision is grossly overestimated (error grossly underestimated) by the ordinary least and weighted least squares regressions based on the individual tree, not on the cluster data. This overestimation of precision increases with the size of the subsample; the larger the number of trees selected from the same cluster, the greater the overestimation. On the other hand, there seems to be a slight underestimation of the precision (overestimation of the error) when the modified least or weighted least squares methods are used. It is possible, however, that this underestimation is due to the fact that with large samples the effect of the finite population correction factor (for the sample clusters that may go as high as 5 percent) has been ignored.

Finally, the two-phase, two-stage sampling procedures we have used, yields unbiased (or at most with negligible bias) estimates and the method to calculate the precision of the estimates is correct, whenever the precision of the first and second phase regressions are properly evaluated.

The conclusions above hold true for the population as constructed here. We feel that our population resembles real world tree populations. The relationship between diameter and height as it appears in real world sample plots is preserved because we have worked with real world plots. The relationship between diameter, height and biomass was, on the other hand, generated by the computer; but this relationship is expected to imitate very closely what happens in real life. Finally, the cluster effect we have added to our population may not be equal in size but it is definitely similar in structure to what we expect in the real world.

Acknowledgments

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

- Arabatzis, A. A.; Cunia, T. Error of biomass regressions: sample trees selected by stratified sampling. In: Proceedings of the workshop on "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.
- Briggs, E. F.; Cunia, T. Effect of cluster sampling in biomass tables construction: linear regression models. *Can. Journal of Forest Research*. 12:255-263; 1982.
- Cunia, T. Some theory on the reliability of volume estimates in a forest inventory sample. *Forest Science*. 11:115-128; 1965.
- Cunia, T. On sampling trees for biomass tables construction: some statistical comments. In: Forest resource inventory workshop proceedings, Vol. 2, W. E. Prayer (Ed.), Colorado State University, Fort Collins, CO; 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. Use of simulation techniques in tree biomass tables construction. In: Proceedings, The 1985 Symposium on system analysis in forest resources. December 8-11, 1985, University of Georgia, Athens, GA; 1985.
- Cunia, T. On the error of biomass estimates in forest inventory; Part 1: its major components. Faculty of Forestry Miscellaneous Publication Number 8 (ESF 85-004), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Cunia, T.; Gillespie, A. J. Cluster sampling and construction of biomass tables: results of a simulation study. In: Proceedings, Third southern biomass energy research conference, March 12-14, 1985, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL; 1985.
- Cunia T.; Michelakackis, J. On the error of tree biomass tables constructed by a two-phase sampling design. *Can. Journal of Forest Research*. 13: 303-313; 1983a.
- 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; 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, NC; 1984.
- Gillespie, A. J.; Cunia, T. Error of biomass regressions: sample trees selected by cluster sampling. In: Proceedings of the workshop on "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.
- 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.
- Michelakackis, J.; Cunia, T. Error of biomass regressions: sample trees selected by double sampling. In: Proceedings of the workshop on "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.

ESTIMATION OF TREE BIOMASS TABLES BY CLUSTER

SAMPLING: RESULTS OF A SIMULATION STUDY

Andrew J. Gillespie and Tiberius Cunia

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

Tree biomass regressions are generally constructed (i) from samples of trees selected by cluster rather than simple random sampling and (ii) by least squares techniques that ignore the cluster effect. The results of a simulation study are reported whereby (i) samples of trees were selected by cluster sampling from a known tree population, (ii) the biomass regression functions were calculated by ordinary least squares methods and by methods modified to take the cluster effect into account and (iii) the estimates of the average biomass per acre were compared to the known true value of the tree population. These results show, among other things, that (i) the estimates of the average biomass per acre based on the ordinary and modified regression techniques are about the same but the estimates of the precision are grossly overstated by the ordinary least squares, and (ii) the bias is not significantly different than zero for all suitably selected models.

Introduction

A common sampling design for taking the biomass inventory of a given forest area is the two phase or double sampling method. Phase one consists of a large sample of trees measured for diameter at breast height, species, and possibly height, along with other variables. These trees are not measured for biomass. The phase two sample is a smaller sample of trees measured for biomass as well as species, diameter, height, and possibly other variables measured in the phase one sample. This sample is used to construct a regression function of biomass on several predictor variables, ordinarily species, diameter, and possibly height. This regression function is then applied to the data of the phase one sample to estimate the mean biomass per tree or mean biomass per unit area.

Simple random sampling of trees is seldom a practical method of sampling a forest. It is more efficient to apply the method of cluster sampling. Cluster sampling involves dividing the forest into overlapping or non-overlapping clusters defined as plots of fixed or variable area. A subset of these clusters is then selected and some or all of the trees from each sample cluster are selected and measured for the variables of interest. The advantage of cluster sampling is the reduced average sampling cost per tree that

is associated with the reduced movement of sampling crews over the forest. The disadvantage is the corresponding reduced amount of information per tree; trees growing close together will tend to be more similar than trees growing farther apart.

One common method of estimating biomass functions from the phase two sample is the method of ordinary least squares (OLS) or weighted least squares (OWLS). These methods assume that the sample trees are selected independently of each other. This assumption is clearly violated in cluster sampling; trees in a given cluster are not selected independently of each other. Cunia (1979) suggested three modifications of the least squares method for cluster sampling. The first modification, using ratio estimator models, has been applied by Kotimaki and Cunia (1981) to two cluster samples of trees. The second modification uses linear regression models, and has been applied to the same two samples by Briggs and Cunia (1982). In both cases, the researchers found that the intraclass correlations had little effect on the point estimates of mean biomass. The estimates of the error of the point estimates, however, were much larger under the modified procedures. The assumptions of the modified models were better satisfied by the actual sampling conditions than were the assumptions of the ordinary models. Therefore, it was assumed that the unmodified procedures underestimated the error and that the modified procedures were better at estimating the error.

The objectives of the present study are to verify and extend the results of Kotimaki and Cunia (1981) and Briggs and Cunia (1982) by examining the accuracy and precision of various biomass function estimation procedures as applied to cluster sampling. We shall only consider the error of the phase two sample; the error of the first phase sample will be ignored. To accomplish this, we shall apply simulation techniques in which a given population of trees is grouped into clusters and repeatedly sampled. All population parameters (including μ = mean biomass per acre) are known exactly. The biomass equation from the phase two sample will be applied to the population stand table, rather than a phase one estimate. The resulting estimation of mean biomass per acre may then be compared directly to the known parameter μ . In the earlier studies listed above, samples and estimates came from natural populations, for which the precise parameters were unknown. Estimates could not be compared to a fixed standard, but rather were analyzed intuitively by how well their inherent assumptions were satisfied.

The simulated population used in this study consists of 22,753 trees from 667 non-empty and 260 empty permanent one-fifth acre sample plots selected from the State of New York. The trees were measured in the field for their species, diameter at breast height d and merchantable height. The total height h and total biomass y (green weight above ground) were simulated for each tree by Monte Carlo techniques described in

detail by Cunia and Michelakackis (1983, 1984a,b) and Cunia, Michelakackis, and Lee (1984). In generating total height and biomass, the techniques took into account the effects of species, diameter, merchantable height, site quality, geographical region, and the intracluster correlation effect, as well as the random effect of the probability distributions of total height and biomass around their respective regression functions.

Sampling Procedures

Our sampling procedure is the following: m clusters ($m = 10, 15, 20, 30, \text{ or } 50$) are chosen at random (without replacement) from the population of 667 non-empty clusters, and p percent ($p = 10, 20, \dots, 100$) of the trees from each cluster are chosen (at random without replacement) and measured for biomass (y), diameter (d), and height (h). The average cluster size is about 34 trees. The study considers only those combinations of m and p which yield an average sample size $(m)(p)(34)$ of at least 10 trees but less than 600 trees. There are 36 such combinations which, for convenience, will be known here as sampling methods. For each of these 36 sampling methods, 100 independent samples of trees were generated by computer from the known population of 22,753 trees.^{1/} This yields a study set of $36 \times 100 = 3600$ different randomly selected cluster samples of trees.

Estimation Models

The main population parameter of interest is μ = average total biomass per acre, defined as the total biomass of all trees in the population divided by the total forest area occupied by the simulated population.^{2/} It is known that μ is equal to 115.549 thousands of pounds per acre.

Two sets of models are initially used to estimate μ . The first set consists of two versions of each of seven ratio estimators, for a total of 14 ratio estimator models. The second set includes four versions of each of five linear functions, for a total of 20 least squares regression models. Initially, we applied each of these 34 estimation procedures to each of the 100 samples consisting of 50 clusters with 30 percent subsampling. After analyzing the results, we selected two ratio estimators and eight least squares regression models for further analysis in application to the remaining 3500 samples of trees. The original 34 models are briefly described below; for more complete descriptions, the reader is referred to Gillespie (1985).

^{1/}This work was done by Sueh Fang Hsu, Graduate student, SUNY-CESF.

^{2/}Total area = (927 plots) $\times$ (1/5 acre per plot) = 185.4 acres.

Ratio Estimators

Ratio estimator models assume that the relationship between tree biomass y and a highly correlated variable x is adequately described by the equation $y = Rx$, where for this study R is defined as μ_y/μ_x , the ratio of the arithmetic means of y and x , and x is defined as either d^2 or d^2h . Note that R is also the ratio of the means per acre or totals over the entire forest of y and x . The purpose of sampling is to estimate R , which is assumed unknown. Models 1 through 7 consist of seven estimation procedures with $x = d^2$; models 8 through 14 consist of the same estimation procedures with $x = d^2h$. These procedures are used to calculate r = the point estimate of R , and S_{rr} = the estimate of σ_{rr} , the variance of r .

Letting first $x = d^2$ and taking the summation over all trees of a given sample, we define the following seven models:

Model 1.

The true regression model of y on x is assumed to be $y = Rx$, with the sample trees selected by simple random sampling and the conditional variance of y given x is constant (homoscedastic):

$$r_1 = \Sigma xy / \Sigma x^2$$

and

$$S_{r_1 r_1} = (\Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2) / (n-1) \Sigma x^2$$

Model 2.

Identical to model 1, but with the conditional variance of y given x assumed to be proportional to x :

$$r_2 = \Sigma y / \Sigma x$$

and

$$S_{r_2 r_2} = (\Sigma (y^2/x) - (\Sigma y)^2 / \Sigma x) / (n-1) \Sigma x$$

Model 3.

Identical to model 2, with the conditional variance of y given x assumed to be proportional to x^2 :

$$r_3 = \Sigma (y/x) / n$$

and

$$S_{r_3 r_3} = (\Sigma (y/x)^2 - (\Sigma (y/x))^2 / n) / (n-1) n$$

Model 4.

Ratio-of-means model which assumes only that the sample trees were collected by simple random sampling:

$$r_4 = \bar{y} / \bar{x}$$

where $\bar{y}$ and $\bar{x}$ are the arithmetic sample means of y and x respectively. This estimator is known to be biased (Cochran, 1977). The variance of r_4 is estimated by

$$S_{r_4 r_4} = \left\{ S_{yy} - 2(r_4)S_{xy} + (r_4)^2 S_{xx} \right\} / n\bar{x}^2$$

where S_{xx} , S_{yy} , and S_{xy} are the usual sample variances and covariance of x and y .

Model 5.

Mean-of-ratios estimator, which also assumes only that the sample trees were collected in a simple random sample:

$$r_5 = \Sigma (y/x) / n = r_3$$

$$S_{r_5 r_5} = (\Sigma (y/x)^2 - (\Sigma (y/x))^2 / n) / n(n-1)$$

This estimator is also known to be biased (Cochran, 1977).

Model 6.

Modified ratio-of-means model, where the trees are assumed collected by a simple random sample of clusters. The modification utilizes the cluster variables

$$u_h = \text{cluster biomass} = \Sigma y \text{ in cluster } h, \\ h = 1, 2, \dots, m,$$

and m = number of clusters in the sample,

and

$$v_h = \Sigma x \text{ in cluster } h$$

with Σ taken over the trees of cluster h . The modified ratio-of-means estimate of R is calculated by

$$r_6 = \bar{u}/\bar{v} = ((\Sigma y)/m) / ((\Sigma x)/m) = r_4 = r_2$$

but the variance has a different formula:

$$S_{r_6 r_6} = \left\{ S_{uu} - 2(r_6)S_{uv} + (r_6)^2 S_{vv} \right\} / m\bar{v}^2$$

where u , v , S_{uu} , S_{vv} , and S_{uv} are the usual sample means, variances, and covariance of u and v .

Model 7.

Modified mean-of-ratios model, where the trees are assumed collected by a simple random sample of clusters. The modification utilizes the cluster variables

$$w_h = \Sigma (y/x) = \text{sum of ratios } (y/x) \text{ in cluster } h.$$

$$n_h = \text{number of sample trees in cluster } h, \text{ and}$$

the modified mean-of-ratios estimates of R and

σ_{rr} are calculated by

$$r_7 = \Sigma w / \Sigma n = \left\{ \Sigma (y/x) \right\} / n = r_5 = r_3$$

$$S_{r_7 r_7} = m \left\{ S_{ww} - 2(r_7)S_{wn} + (r_7)^2 S_{nn} \right\} / n^2$$

where S_{ww} , S_{nn} , and S_{wn} are the usual sample variances and covariance of w and n . Although $r_2 = r_4 = r_6$, the variances and assumptions of each model are distinct; therefore these are all different models. Similarly, while $r_3 = r_5 = r_7$ and $S_{r_3 r_3} = S_{r_5 r_5} = S_{r_7 r_7}$, the assumptions of these three models are also distinct, hence they represent three different models.

Models 8 - 14 consist of the same seven estimators with $x = d^2 h$. Once r and S_{rr} have been estimated, we calculate

$z_i = r_i \mu_x$ = the point estimate of mean biomass per acre, and

$V_i = S_{r_i r_i} (\mu_x)^2$ = the estimate of the variance of z_i ,

where μ_x = the known average "sum of x " per acre

Regression Estimators

The study considers 20 procedures (models) for estimating linear regressions equations of y as a function of d or d and h . These models consist of 20 combinations of five linear regression functions with four least squares estimation approaches. The five linear functions considered are:

1. $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$
2. $\hat{y} = \beta_1 + \beta_3 d^2$
3. $\hat{y} = \beta_1 + \beta_4 d^2 h$
4. $\hat{y} = \beta_4 d^2 h$
5. $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2 + \beta_4 d^2 h + \beta_5 d h + \beta_6 h$

with function 5 retaining only those terms for which the estimated coefficients are statistically discernible from 0. These functions are expressed in the standard mathematical notation $\hat{y} = [\beta]'[x]$, where $[\beta]$ = the vector of regression coefficients, $[\beta]'$ denotes the transposed vector $[\beta]$, and $[x]$ denotes the vector of predictor variables corresponding to $[\beta]$.

We consider four approaches to calculate $[\beta]$, the point estimate of the vector $[\beta]$ of regression coefficients. The first approach is that of ordinary least squares (OLS), applied to the individual trees of the sample with the conditional variance of y given $[x]$ assumed to be constant. The second approach is that of ordinary weighted least squares (OWLS), also applied to the individual trees of the sample, but with the conditional variance of y given $[x]$ assumed to be proportional to (i) d^4 for models which use d alone or (ii) $d^4 h^2$ for models which use

both d and h as independent variables. These first two approaches are both well documented in statistical literature as, for example, by Draper and Smith (1981), and by Neter, Wasserman, and Kutner (1983).

Approaches 3 and 4 are modified regression techniques which utilize cluster variables, rather than individual tree values. Since most forest inventories utilize cluster samples (rather than simple random samples of trees), modified techniques may be more accurate and precise than standard techniques for which the assumptions are not satisfied. These modifications are briefly described below. For more detailed descriptions of the rationale and method of modification for cluster sampling, the reader is referred to Cunia (1979, 1981) and Briggs and Cunia (1982). Modification for approaches 3 and 4 consists of creating m new cluster variables u_h and v_h by summing y_i and x_i over each cluster h ($i = 1, 2, \dots, n_h$ and $h = 1, 2, \dots, m$, where n_h = number of trees in cluster h):

$$u_h = \sum y_i \text{ in cluster } h, \text{ and}$$

$$v_h = \sum x_i \text{ in cluster } h.$$

The new sample size will now be the number of non-empty clusters in the sample. The new model is now assumed to be $\hat{u} = [\beta]'[v]$. Comparison of approaches 1 and 2 versus approaches 3 and 4 will illustrate the error made by treating a cluster sample as if it had been collected by simple random sampling.

We assume that the variables y are normally distributed and selected independently within any given cluster, that the conditional variance of y given [x] is equal to $a_i^2 \sigma^2$, where a_i^2 is known for each tree and σ^2 is an unknown constant, and that the other assumptions of least squares methods are satisfied. Since u is defined as the summation of several random variables y in a cluster, it is reasonable to further assume that the conditional variance of u given [v] is proportional to $(\sum a_i^2) \sigma^2$. Approach 3 assumes that the variables u and v of the various clusters are statistically independent, and that $a_i^2 = 1$ so that the conditional variance of u given [v] is proportional to $\sum (1)^2 = n_h$, the number of sample trees in cluster h. This is the modified least squares (MLS) approach. Approach 4 makes the same assumptions except that a_i^2 is assumed to be d^4 (functions 1,2) or $d^4 n^{21}$ (functions 3,4,5). The conditional variance of u given [v] is assumed to be proportional to $\sum d^4$ or $\sum d^4 n^2$. This is the modified weighted least squares (MWLS) approach.

To summarize, regression models 15-19 consist of the 5 linear functions estimated by approach 1; models 20-24 are the 5 linear functions estimated by approach 2; models 25-29 are the 5 linear functions estimated by approach 3; and models 30-34 are the 5 linear functions estimated by approach 4.

Once we have obtained the estimate [b] of $[\beta]$ and $[S_{bb}]$ of $[\sigma_{bb}]$ = the covariance matrix of

[b], the estimators z_i and V_i for regression model i are given by:

$z_i = [b_i]'[\mu_i]$ = the point estimate of mean biomass per acre, and

$V_i = [\mu_i]'[S_{bb}][\mu_i]$ = the estimate of the variance of z_i ,

where $[\mu_i]$ = a column vector of "means per acre" of each of the respective predictor variables x included in model i. These means are all known exactly for the simulated population.

We point out that, in terms of calculations (but not assumptions), (i) models 3 and 5 are identical, (ii) models 8 and 18 are identical, and (iii) models 10, 12, and 23 are identical. We therefore consider 30 rather than 34 distinct estimation procedures.

Analysis Procedure

For a given estimation procedure i ($i = 1, 2, \dots, 34$) and each sample j from a given sampling method ($j = 1, 2, \dots, 100$), we calculate the sample statistic z_{ij} and sample variance V_{ij} which estimate the population mean μ and variance σ_{zz} respectively. We also note whether μ is above, within, or below the 95 percent and 99 percent confidence intervals of z_{ij} . Finally, we summarize all 100 estimates z_{ij} and V_{ij} for each model-sampling method combination with the statistics:

1. $\bar{z}_i = \sum z_{ij} / 100$
2. $S_{zz} = (\sum z_{ij}^2 - (\sum z_{ij})^2 / 100) / 99$
3. $\bar{V}_i = \sum V_{ij} / 100$
4. $(\bar{z}_i - \mu)$
5. $t_i = (\bar{z}_i - \mu) / \sqrt{S_{zz} / 100}$
6. The number of times (out of 100 trials) that μ was above, within, or below the 95 and 99 percent confidence intervals of z_{ij} .

$\bar{V}_i$ is an average estimate of the variance of z_{ij} under the assumptions of the model i. If the model assumptions are strictly satisfied, then each V_{ij} and their average $\bar{V}_i$ are unbiased estimators of the variance σ_{zz} of z_{ij} . Because the 100 estimates z_{ij} are generated by the same random process (sampling method and estimation procedure), they may be considered to be 100 statistically independent random variables. In this case, unbiased estimates of μ and σ_{zz} are given by the usual statistics $\bar{z}$ and S_{zz} .

The analysis focuses on the variation of these summary statistics between models for a given sampling method, and within each given model over varying sampling procedures.

Analysis Part I

For the initial analysis, we apply the 34 models to 100 samples from the 50 cluster, 30 percent subsampling method. The results of this analysis are given in Cunia and Gillespie (1985), and are briefly summarized below.

The bias of the ratio-of-means estimator (models 2, 4, 6, 9, 11, and 13) is small and generally not significantly different than zero for our sample data. This is true for both $x = d^2$ and $x = d^2h$. The other ratio estimators (models 1, 3, 5, 7, 8, 10, 12, 14) have statistically significant bias that may vary from 2 to 7 percent of the mean μ . The estimate S_{zz} of the variance of z is very high for models 1 and 8, and more reasonable for the other models. The estimate V of the variance of z is (i) much lower than S_{zz} for models which ignore the cluster effect, and (ii) slightly higher than S_{zz} for models which allow for the cluster effect. Hence the only acceptable ratio estimator models are models 6 and 13, ratio-of-means estimators which account for the effects of cluster sampling. The adjustment for cluster sampling does not affect the accuracy of the estimation procedure. It does affect the estimated precision of the point estimate; the average estimated precision $\bar{V}$ for modified models 6 and 13 is approximately three times as large as $\bar{V}$ of the corresponding unmodified models 4 and 11. This agrees with the findings of Kotimaki and Cunia (1981), and is illustrated by the behavior of the confidence interval; only the confidence intervals for models 6 and 13 included μ an acceptable number of times out of 100 replications.

For the linear regression models 15 to 34, this first analysis indicates that (i) regression models 18, 23, 28, and 33 (with the equation $y = \beta_4 d^2 h$) are very erratic, with large bias and low precision; (ii) the bias of the other 16 regression estimators is generally small and not statistically different from zero; (iii) the models which ignore the effect of cluster sampling tend to grossly overestimate the precision of the point estimators, with the estimate V much lower than the corresponding S_{zz} , and (iv) modified models are much better at estimating the precision of z_{ij} , as illustrated by the relatively small difference between $\bar{V}$ and S_{zz} , as well as the degree of confidence interval reliability. This agrees so far with the findings of Briggs and Cunia (1982).

At this point, the field of models is narrowed down to two ratio estimators (models 6 and 13) and 8 regression estimators (models 15, 17, 20, 22, 25, 27, 30, and 32) for complete analysis over the remaining 35 sampling procedures. These models are selected because they all yielded generally unbiased estimates of μ but different estimates of the error. Part two of the analysis will look at the behavior of these models over varying sampling methods.

Analysis Part II

Analysis of estimated bias ($\bar{z} - \mu$).

The size of the estimated bias ($\bar{z} - \mu$) varies within and between the 36 sampling methods. A set of tables and graphs given by Gillespie (1985) shows this variation. Following are the main conclusions that we have derived from the analysis of these tables and graphs.

Within the models using d alone, none of the estimates of bias are significantly different from zero at the 95 percent confidence level. The conclusions are the same for the models using both d and h except for the 10 (and sometimes 15) cluster samples, for which the bias is significant. There seems to be no reason for this inconsistency. We have concluded that, in general, the bias of z is small and, for our data, not significantly different than zero.

It is of interest to mention also that (i) for models using d alone, the absolute value of the bias seems to remain approximately constant over samples of any size (except for the 10 or 15 cluster samples, where it is larger), and (ii) there is a tendency for the bias of the models with d and h to decrease when, for a fixed number of sample trees, the number of sample clusters increases. This last phenomenon is assumed to be due to random effects.

Analysis of $\sqrt{S_{zz}}$ the estimate of $\sqrt{\sigma_{zz}}$.

The statistic $\sqrt{S_{zz}}$ is defined as the estimate of the standard deviation (error) of z . For a given model-sampling method i , it is calculated from the 100 estimates z_{ij} , and is thus independent of any model assumptions concerning variance estimation. Table 1 lists $\sqrt{S_{zz}}$ for all 360 model-sampling method combinations.

As expected, $\sqrt{S_{zz}}$ decreases as average sample size increases and as the number of clusters increases for a fixed sample size. As percent subsampling increases for a fixed number of clusters, $\sqrt{S_{zz}}$ decreases until approximately 30 to 40 percent of trees on a cluster have been sampled, then remains approximately constant as percent subsampling increases. Models using d alone have consistently higher $\sqrt{S_{zz}}$ (lower precision) than similar models using both d and h ; this is due to the additional information contributed by h in the estimation of biomass. Models using d alone show greater improvement in precision due to increases in the number of clusters sampled, with the difference in the precision estimates of these two groups becoming negligible when 50 clusters are sampled.

The OLS, OWLS, and ratio estimator models which use the same independent variables have roughly equal standard errors for any sampling method. The same relationship exists between MLS and MWLS models. This shows that, with respect to the point estimation of μ , unweighted regression procedures are very robust; we do not get

Table 1 The estimate $\sqrt{S_{zz}}$ (thousands of pounds per acre) of the standard deviation of z , by model and sampling method; m = number of clusters, p = percent of trees sampled per cluster.

Sampling Method		Model									
m	p	6	13	15	17	20	22	25	27	30	32
10	10	9.72	6.63	10.31	6.74	9.63	6.58	11.82	7.70	11.05	7.54
	20	8.41	5.96	8.48	6.07	8.52	5.62	9.14	6.06	8.57	6.04
	30	7.96	5.58	7.85	5.71	7.98	5.32	8.64	6.12	8.40	6.13
	40	7.71	5.46	7.51	5.54	7.63	5.27	8.09	5.90	8.23	6.01
	50	7.71	5.30	7.42	5.41	7.52	5.14	8.11	5.90	8.24	6.02
	60	7.54	5.16	7.36	5.22	7.40	5.01	7.85	5.78	7.84	5.82
	70	7.52	5.13	7.33	5.18	7.42	5.04	8.04	5.63	8.04	5.65
	80	7.57	5.13	7.42	5.17	7.46	5.00	8.39	5.69	8.36	5.56
	90	7.51	4.99	7.44	5.05	7.39	4.90	8.55	5.57	8.45	5.41
	100	7.53	4.90	7.42	4.97	7.37	4.86	8.55	5.44	8.26	5.29
15	10	7.26	4.91	7.45	5.47	6.94	5.43	8.03	5.87	7.52	5.61
	20	6.55	4.76	6.73	4.97	6.56	4.94	7.34	5.32	6.77	5.28
	30	6.29	4.68	6.31	4.80	6.40	4.83	6.59	4.94	6.25	5.06
	40	6.18	4.71	6.22	4.83	6.16	4.82	6.79	4.98	6.51	4.95
	50	6.15	4.62	6.18	4.78	6.12	4.71	6.57	5.00	6.46	4.97
	60	6.02	4.47	5.99	4.62	6.00	4.60	6.41	4.93	6.32	4.86
	70	6.01	4.44	6.01	4.58	6.05	4.56	6.43	4.92	6.41	4.82
	80	6.00	4.45	6.01	4.57	6.01	4.55	6.35	4.88	6.54	4.82
	90	6.06	4.35	6.07	4.46	6.06	4.49	6.47	4.76	6.63	4.74
	100	6.09	4.33	6.05	4.49	6.05	4.54	6.47	4.78	6.64	4.68
20	10	5.62	4.09	5.59	4.32	5.59	4.38	5.79	4.67	5.65	4.30
	20	5.21	3.60	4.91	3.60	5.03	3.82	5.36	3.68	5.14	4.06
	30	4.95	3.62	5.06	3.57	5.03	3.69	5.47	3.64	5.19	3.77
	40	4.87	3.61	4.90	3.56	4.80	3.71	5.37	3.70	5.13	3.83
	50	4.76	3.50	4.75	3.46	4.66	3.56	5.19	3.71	5.00	3.78
	60	4.77	3.45	4.71	3.42	4.64	3.47	5.15	3.69	4.99	3.78
	70	4.74	3.36	4.71	3.36	4.62	3.41	5.01	3.63	5.00	3.69
	80	4.74	3.36	4.83	3.35	4.70	3.47	4.95	3.60	5.07	3.65
30	10	5.08	3.56	5.21	3.87	4.88	3.44	5.25	4.07	5.14	3.73
	20	4.58	3.16	4.49	3.21	4.37	3.24	4.60	3.37	4.55	3.53
	30	4.35	3.17	4.33	3.19	4.27	3.34	4.48	3.28	4.47	3.34
	40	4.21	3.17	4.05	3.18	4.02	3.30	4.07	3.24	4.09	3.38
	50	4.00	3.02	3.99	3.02	3.93	3.16	4.06	3.12	3.88	3.21
50	10	4.20	2.79	3.94	2.85	3.86	2.85	3.91	2.97	3.83	3.08
	20	3.68	2.56	3.49	2.52	3.53	2.47	3.62	2.60	3.43	2.86
	30	3.31	2.49	3.26	2.46	3.33	2.50	3.32	2.52	3.22	2.61

much more precise point estimates by using the slightly more complicated weighted procedures. However, $\sqrt{S_{zz}}$ is generally slightly higher for modified (MLS or MWLS) versions of a model than for unmodified (OLS or OWLS) versions of the same model. In other words, modified models are slightly less precise than analogous unmodified models. This difference is most noticeable with 10 cluster samples, and decreases as the number of clusters sampled increases. Although the difference in most cases is not statistically discernible (Gillespie 1985), it is nonetheless interesting that this trend is so consistent. One possible explanation for this trend is that the range of individual tree measurements y , d , and h is relatively larger than the range of average summed cluster variables u and v . This trend did not occur in earlier studies by Briggs

and Cunia (1982) and Kotimaki and Cunia (1981).

Analysis of $\sqrt{\bar{V}}$, the estimate of $\sqrt{\sigma_{zz}}$

The statistic $\bar{V}$ is the estimate of σ_{zz} as calculated under the assumptions of each given model. We will focus on the standard deviation $\sqrt{\bar{V}}$, which is calculated as the square root of the arithmetic average of the 100 estimates V_{ij} of the variance of z_{ij} (one model applied to 100 samples). Table 2 shows $\sqrt{\bar{V}}$ for all 360 model-sampling method combinations in the study.

The estimate $\sqrt{\bar{V}}$ decreases as average sample size increases. For the unmodified regression models, $\sqrt{\bar{V}}$ seems directly dependent upon the number of trees in the sample size, regardless

Table 2 The estimate $\sqrt{V}$ (thousands of pounds per acre) of the standard deviation of z , by model and sampling method. m = number of clusters, p = percent of trees sampled per cluster.

Sampling Method		Model									
m	p	6	13	15	17	20	22	25	27	30	32
10	10	9.66	6.46	7.35	5.05	8.29	6.19	10.31	6.13	12.51	7.80
	20	8.45	5.96	5.26	3.80	5.60	4.31	9.45	5.63	11.27	6.78
	30	7.97	5.76	4.50	3.21	4.51	3.46	8.85	5.43	9.88	6.21
	40	7.64	5.49	3.89	2.80	3.87	2.97	8.56	5.19	9.55	6.05
	50	7.54	5.52	3.45	2.52	3.44	2.66	8.13	5.33	9.36	6.04
	60	7.39	5.46	3.12	2.29	3.14	2.42	7.98	5.20	9.21	6.10
	70	7.38	5.38	2.90	2.13	2.89	2.23	7.94	5.30	9.02	6.06
	80	7.23	5.32	2.69	1.99	2.71	2.08	7.77	5.16	8.82	6.00
	90	7.18	5.29	2.55	1.89	2.55	1.96	7.95	5.13	8.90	5.96
	100	7.16	5.27	2.42	1.81	2.41	1.86	7.87	5.13	8.82	5.87
15	10	7.92	5.54	5.92	4.29	6.38	4.82	7.84	5.58	9.32	6.32
	20	6.76	4.92	4.33	3.17	4.34	3.36	7.21	5.01	7.95	5.43
	30	6.36	4.72	3.63	2.65	3.54	2.72	6.71	4.71	7.26	5.04
	40	6.10	4.55	3.12	2.30	3.05	2.35	6.37	4.48	6.95	4.85
	50	6.03	4.48	2.79	2.05	2.71	2.10	6.19	4.38	6.80	4.77
	60	5.93	4.44	2.53	1.87	2.48	1.92	6.05	4.29	6.71	4.74
	70	5.93	4.37	2.35	1.73	2.30	1.77	6.03	4.25	6.65	4.70
	80	5.83	4.34	2.18	1.63	2.15	1.66	5.86	4.20	6.53	4.68
	90	5.77	4.29	2.06	1.54	2.03	1.56	5.84	4.13	6.51	4.64
	100	5.75	4.28	1.97	1.48	1.91	1.48	5.81	4.13	6.44	4.58
20	10	6.85	4.92	5.16	3.77	5.49	4.15	6.68	4.86	7.71	5.45
	20	5.90	4.27	3.79	2.77	3.76	2.88	5.99	4.31	6.58	4.67
	30	5.59	4.07	3.17	2.31	3.08	2.35	5.73	4.09	6.20	4.39
	40	5.38	3.91	2.73	2.01	2.65	2.04	5.52	3.91	5.95	4.23
	50	5.30	3.86	2.44	1.79	2.36	1.82	5.35	3.85	5.81	4.16
	60	5.20	3.81	2.21	1.63	2.16	1.67	5.22	3.77	5.72	4.11
	70	5.17	3.77	2.04	1.51	2.01	1.54	5.16	3.72	5.67	4.07
	80	5.08	3.74	1.92	1.43	1.87	1.45	5.00	3.67	5.57	4.05
30	10	5.63	4.04	4.29	3.18	4.32	3.36	5.37	4.01	5.89	4.39
	20	4.84	3.48	3.11	2.32	3.02	2.34	4.70	3.50	5.07	3.74
	30	4.54	3.27	2.58	1.90	2.47	1.92	4.41	3.27	4.77	3.52
	40	4.38	3.16	2.23	1.66	2.14	1.66	4.27	3.15	4.63	3.41
	50	4.30	3.10	1.99	1.48	1.91	1.49	4.15	3.09	4.55	3.34
50	10	4.41	3.20	3.44	2.58	3.33	2.61	4.32	3.25	4.52	3.41
	20	3.80	2.73	2.45	1.82	2.35	1.83	3.70	2.77	3.92	2.93
	30	3.57	2.56	2.04	1.49	1.91	1.50	3.46	2.57	3.69	2.76

of the sampling method. For the ratio estimators and modified regression models, all of which account for the cluster effect of sampling, $\sqrt{V}$ behaves much like $\sqrt{S_{zz}}$; specifically, $\sqrt{V}$ decreases with an increase in the number of clusters sampled, as well as with increased value of p (up to 30 to 40 percent). For these models, the effect of increasing the number of clusters does more to reduce $\sqrt{V}$ than increasing percent subsampling. Estimates of $\sqrt{V}$ from models which use only d are generally 1.3 to 1.5 times as large as estimates from models which use d and h .

With small sample sizes, $\sqrt{V}$ as estimated by weighted least squares models is generally slightly higher (10 to 30 percent) than $\sqrt{V}$ as

estimated by unweighted least squares regression models for similar sampling methods. This difference gets smaller as the number of clusters and/or the percent subsampling increases. The estimates of $\sqrt{V}$ from modified regression models are 30 to 50 percent higher than estimates from unmodified regression models. The estimates of $\sqrt{V}$ from ratio estimator models are larger than those from unmodified regressions, but smaller than those from modified regressions (which use the same independent variable(s)).

Table 3 The ratio $\sqrt{\bar{V}/S_{zz}}$ of the two different estimates of the standard deviation of z , by model and sampling method. m = number of clusters sampled, p = percent of trees sampled per cluster.

Sampling Method		Model									
m	p	6	13	15	17	20	22	25	27	30	32
10	10	0.99	0.97	0.71	0.75	0.86	0.94	0.87	0.80	1.13	1.04
	20	1.00	1.00	0.62	0.63	0.66	0.77	1.03	0.93	1.32	1.12
	30	1.00	1.03	0.57	0.56	0.57	0.65	1.02	0.89	1.18	1.01
	40	0.99	1.01	0.52	0.51	0.51	0.56	1.06	0.88	1.16	1.01
	50	0.98	1.04	0.46	0.47	0.46	0.52	1.00	0.90	1.14	1.00
	60	0.98	1.06	0.42	0.44	0.42	0.48	1.02	0.90	1.18	1.05
	70	0.98	1.05	0.40	0.41	0.39	0.44	0.99	0.94	1.12	1.07
	80	0.96	1.04	0.36	0.38	0.36	0.42	0.93	0.91	1.06	1.08
	90	0.96	1.06	0.34	0.37	0.35	0.40	0.93	0.92	1.05	1.10
	100	0.95	1.07	0.33	0.36	0.33	0.38	0.92	0.94	1.07	1.11
15	10	1.09	1.13	0.79	0.78	0.92	0.89	0.98	0.95	1.24	1.13
	20	1.03	1.03	0.64	0.64	0.66	0.68	0.98	0.94	1.17	1.03
	30	1.01	1.01	0.57	0.55	0.55	0.56	1.02	0.95	1.16	1.00
	40	0.99	0.96	0.50	0.48	0.50	0.49	0.94	0.90	1.07	0.98
	50	0.98	0.97	0.45	0.43	0.44	0.45	0.94	0.88	1.05	0.96
	60	0.99	0.99	0.42	0.40	0.41	0.42	0.94	0.87	1.06	0.97
	70	0.99	0.98	0.39	0.38	0.38	0.39	0.94	0.86	1.04	0.97
	80	0.97	0.98	0.36	0.36	0.36	0.36	0.92	0.86	1.00	0.97
	90	0.95	0.99	0.34	0.35	0.33	0.35	0.90	0.87	0.98	0.98
	100	0.94	0.99	0.33	0.33	0.32	0.33	0.90	0.86	0.97	0.98
20	10	1.22	1.20	0.92	0.87	0.98	0.95	1.15	1.04	1.37	1.27
	20	1.13	1.19	0.77	0.77	0.75	0.75	1.12	1.17	1.28	1.15
	30	1.13	1.12	0.63	0.65	0.61	0.64	1.05	1.12	1.19	1.16
	40	1.10	1.08	0.56	0.56	0.55	0.55	1.03	1.06	1.16	1.11
	50	1.11	1.10	0.51	0.52	0.51	0.51	1.03	1.04	1.16	1.10
	60	1.09	1.10	0.47	0.48	0.47	0.48	1.01	1.02	1.15	1.09
	70	1.09	1.12	0.43	0.45	0.43	0.45	1.03	1.02	1.13	1.10
	80	1.07	1.11	0.40	0.43	0.40	0.42	1.01	1.02	1.10	1.11
30	10	1.11	1.13	0.82	0.82	0.89	0.98	1.02	0.98	1.15	1.18
	20	1.06	1.10	0.69	0.72	0.69	0.72	1.02	1.04	1.12	1.06
	30	1.04	1.03	0.59	0.60	0.58	0.57	0.99	1.00	1.07	1.05
	40	1.04	1.00	0.55	0.52	0.53	0.50	1.05	0.97	1.13	1.01
	50	1.08	1.03	0.50	0.49	0.49	0.47	1.02	0.99	1.17	1.04
50	10	1.05	1.15	0.87	0.91	0.86	0.92	1.10	1.10	1.18	1.11
	20	1.03	1.07	0.70	0.72	0.66	0.74	1.02	1.06	1.14	1.02
	30	1.08	1.03	0.62	0.60	0.57	0.60	1.04	1.02	1.14	1.06

Analysis of the ratio $\sqrt{\bar{V}/S_{zz}}$

The ratio $\sqrt{\bar{V}/S_{zz}}$ illustrates the difference between the assumed and the actual precision of the various models. Table 3 lists this ratio for all 360 model-sampling method combinations.

For the ratio of mean estimators and modified regression, the ratio $\sqrt{\bar{V}/S_{zz}}$ is approximately 1.0 or slightly higher, staying constant regardless of changes in the sampling method or sample size. For unmodified regression models, $\sqrt{\bar{V}/S_{zz}}$ is generally between .4 and 1.0, decreasing with increases in percent subsampling from a fixed number of clusters, and increasing with an

increase in the number of clusters sampled for a fixed percentage of trees. The net effect is a slight decrease in the ratio as the average sample size increases, indicating that $\bar{V}$ decreases relative to S_{zz} .

In general, the ratio $\sqrt{\bar{V}/S_{zz}}$ is larger for weighted regression models than for similar but unweighted regression models. The ratio is also much larger for modified regression models (approximately 1.0 or higher) than for similar but unmodified regression models (from .4 to 1.0).

In all cases of unmodified regression models, the ratio is less than 1.0. If S_{zz} is

accepted as a better (and unbiased) estimate of σ_{zz} , then V underestimates σ_{zz} in unmodified models. This is expected, since the assumptions of independence among sample elements is violated.

These relationships are emphasized by the analysis of confidence interval reliability. Unmodified regression models yield 95 percent confidence intervals which were only correct 40 percent of the time. Ratio estimators and modified unweighted regression models yield 95 percent confidence intervals which are correct approximately 95 percent of the time (except when samples came from 10 clusters). Modified weighted regression models yield 95 percent confidence intervals which are 95 to 100 percent correct, implying that they may be too large and that the weighted modified procedure overestimates the error of the estimate.

It is important to note that the study population was simulated by a weighted, unmodified regression model with conditional variance set proportional to $d^4 h^2$. In the modified weighted regression models, the cluster biomass was defined as $u = \sum y$, with the variance of u assumed to be proportional to the sum of the a_i^2 (with $a_i^2 = d^4$ or $d^4 h^2$). A better assumption (based on the success of the modified unweighted regression models) seems to be that the conditional variance of u is proportional to the number of sample trees in the cluster, n_h , even when the individual tree biomass conditional variance is known to be proportional to $d^4 h^2$ (Gillespie, 1985).

Summary Comments

The main objectives of this study were to verify the results of Briggs and Cunia (1982) and Kotimaki and Cunia (1981) concerning the estimation of biomass tables by cluster sampling. These earlier studies found that the point estimates of biomass were approximately equally accurate whether or not the estimation procedure considered the cluster effect of sampling. However, they felt (but could not prove) that ignoring the cluster effect caused a large underestimation of the error of such point estimates.

Our conclusions generally agree with these results. The ratio-of-means estimator and parabolic least squares regression models were practically unbiased estimators of μ , especially when samples were selected from at least 15 clusters. There was no statistically discernible difference between modified and unmodified procedures in calculating point estimates of biomass.

The estimate S_{zz} of error of the point estimates was (i) much lower for models using both d and h than for models using d alone; (ii) roughly equal for weighted and unweighted versions of a similar regression model; and (iii) slightly higher for modified (for cluster effect) versions of regression models than for unmodified versions of the same model.

For all models, S_{zz} decreased with both an increase in the number of clusters sampled for a fixed average sample size, and with increased percentage p subsampling from a fixed number of clusters. This decrease tapers off after 30 to 40 percent of subsampling.

The estimate V of the error of the point estimates reflects the model assumptions concerning variance estimation. For unmodified regression models, V was consistently less than S_{zz} , which implies that ignoring the cluster effect will lead to underestimation of the error, or overestimation of precision. For modified models and the ratio of means estimators which considered the cluster effect, V was approximately equal to or slightly greater than S_{zz} , except for modified weighted regression models when V was consistently larger (by up to 25 percent) than S_{zz} . This overestimation of the true error by modified weighted regression models was not reported in the earlier studies. The consistency of this behavior over the variety of sampling methods studied implies that the assumptions of the modified weighted regression estimators concerning variance estimation may not be correct for our study population. This was emphasized by the behavior of confidence interval reliability, where counts of confidence intervals which included the known population mean μ (out of sets of 100 trials) were consistently greater than expected. Unweighted modified regression models yielded the best estimates of error, implying that the conditional variance of the cluster sum of biomass is proportional to the number of trees in the cluster, rather than the sum of d^4 or $d^4 h^2$.

The present study was limited to simple random samples of clusters and of a fixed proportion of trees within clusters. Future studies may have to consider (i) selecting trees from clusters with probability proportional to some measure of tree size; (ii) consideration of other estimation procedures, such as generalized least squares or a hybrid regression which combines OLS point estimates and MLS error estimates; and (iii) estimation techniques applied to portions of samples, for example trees above a specified DBH.

Acknowledgements

This paper is based on the Master's thesis of the senior author. The research was funded by the Research Foundation of the State of New York, the United States Department of Agriculture-Forest Service, and the Department of Energy, grant number 23-524. We thank Ms. Sueh-Fang Hsu for permitting use of her simulated samples and allowing modification of one of her computer programs.

Literature Cited

- Briggs, E. F.; Cunia, T. Effect of cluster sampling in biomass table construction: linear regression models. *Canadian Journal of Forest Research* 12: 255-263; 1982.
- Cochran, W. G. Sampling techniques. 3d ed. New York: John Wiley & Sons; 1977. 428 p.
- Cunia, T. On tree biomass tables and regression: some statistical comments. In: 1979 Forest resource inventories workshop proceedings, W. E. Frayer, (Ed.) Colorado State University, Fort Collins, CO; 1979: v. II, 629-642.
- Cunia, T.; Gillespie, A. J. Cluster sampling and construction of biomass tables: Results of a simulation study. In: Proceedings, third annual southern forest biomass workshop, 1985 March 12-14; University of Florida, Gainesville, FL.
- Cunia, T.; Michelakackis, J. A method to construct a forest biomass population model. In: Proceedings, Resources inventories for monitoring changes and trends; 1983; J. F. Bell and T. Atterbury, (Eds.); Oregon State University Corvallis, OR; 1983: 558-562.
- Cunia, T.; Michelakackis, J. A Monte Carlo technique for generating total heights of forest trees. Syracuse, NY: SUNY College of Environmental Science and Forestry; 1984a; School of Forestry Miscellaneous Publication Number 4 (ESF 84-018). 24 p.
- Cunia, T.; Michelakackis, J. Constructing forest biomass populations for simulated sampling. Syracuse, NY: SUNY College of Environmental Science and Forestry; 1984b; School of Forestry Miscellaneous Publication Number 5 (ESF 84-019). 46 p.
- Cunia, T.; Michelakackis, J.; Lee, S. Generating total tree heights by a Monte Carlo technique. In: Proceedings of the Fifth Annual Southern Forest Biomass Workshop; 1983 June 15-17; Charleston, SC. R. F. Daniels and P. H. Dunham (Eds); USDA Forest Service, Southeastern Forest Experiment Station, Asheville, NC; 1984.
- Draper, N. R.; Smith, H. Applied regression analysis. 2d ed. New York: John Wiley and Sons; 1981.
- Gillespie, Andrew J.R. Estimation of biomass tables by cluster sampling: a simulation study. M.S. thesis, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1985. 302 p.
- Kotimaki, T. A.; Cunia, T. Effect of cluster sampling in biomass table construction: ratio estimators models. *Canadian Journal of Forest Research* 11: 475-486; 1981.
- Neter, J., W.; Wasserman, W.; Kutner, M. H. Applied linear regression models. Homewood, IL: Richard D. Irwin, Inc; 1983. 547 p.

ERROR OF BIOMASS REGRESSIONS: SAMPLE TREES

SELECTED BY STRATIFIED SAMPLING

Alexandros Arabatzis and Tiberius Cunia

Graduate student and Professor of Statistics and Operations Research respectively, SUNY College of Environmental Science and Forestry, Syracuse, NY 13210

It is common to select sample trees for biomass tables construction by stratified, two-stage cluster sampling whereby (i) several clusters (plots) of trees are selected at random from each of a given number of strata and (ii) from each sample cluster a certain percent of trees is selected at random to be measured for biomass, diameter and height. The common procedure to estimate a biomass regression function applicable to the entire forest area (containing all strata) is to use the least or weighted least squares method applied to the individual tree data irrespective of the stratum from which each tree has been selected. Simulation techniques were used to evaluate the validity of the inferences made when (i) the above procedure is used to estimate the biomass regression function and the average biomass per unit area of a given forest, or (ii) this procedure is replaced by alternative procedures that take into account the effect of the stratum from which the individual trees are being selected. It is shown that (i) the common estimation procedure is generally poor as it may lead to biased results and (ii) the proper estimation procedure is to calculate biomass regression functions separately by stratum.

Introduction

Cunia (1986) describes a computer simulation process whereby (i) samples of trees are selected by a variety of sampling techniques from a large population of forest trees, (ii) for each individual sample so selected, a variety of statistical procedures are used to estimate the regression function of tree biomass on diameter or diameter and height, (iii) each biomass regression of each individual sample is applied to the tree population values of diameter and height to estimate the mean biomass per acre and (iv) by comparing the estimates with the known value of the population mean biomass per acre, inferences are made about the probability behavior of the estimates; inferences about the effect, if any, of the sampling method and estimation procedure on the bias of the estimates, their precision and their estimated precision.

He has described three major sampling techniques. The first is the two-stage random sampling where m sample plots are selected from the population of 667 non-empty plots (clusters of trees) by simple random sampling without replace-

ment and the trees from the selected plots are subsampled by a variety of sampling procedures. A second sampling method, defined as two-phase, two-stage random sampling, generates samples consisting of two, two-stage random samples; a first phase, relatively small sample where the trees are measured for diameter, height and biomass and a second phase, relatively large sample with the trees measured for diameter and height but not biomass. The third sampling method, denoted here as stratified, two-stage sampling is similar to the first; the difference is that the m sample plots of the first stage are selected by stratified, not simple random sampling.

Cunia and Gillespie (1985), Gillespie (1985), Gillespie and Cunia (1986) and Michalakakis and Cunia (1985, 1986) discuss in more detail the simulation processes and the results obtained by the first two sampling methods. In the present paper we shall discuss the results obtained by the third method. The detailed results are reported by Arabatzis (1986). More specifically, we shall describe the simulation process applied to stratified two-stage sampling and its five components, namely, the population being simulated, the sampling method, the estimation procedures for the biomass regression function, the application of the biomass regression to estimate the mean biomass per acre and finally, the analysis of the results obtained. Not to repeat processes already described or conclusions already reached, we shall refer heavily to the above-mentioned papers.

Forest Tree Populations Being Simulated

The basic tree population, hereby called Population 1 has been constructed by a procedure described in a series of papers by Cunia and Michelakakis (1983, 1984a,b) and Cunia, Michelakakis and Lee (1984) and summarized in a paper by Cunia (1986). For more details the interested reader is referred to these papers. It suffices to state here that (i) the basic data used to construct the tree population consists of actual field measurements of diameter at breast height (d), merchantable height (h_m) and species, among other things, performed on 22753 trees of merchantable size ($d > 5$ inches) contained in 927 one-fifth acre plots (260 of which are empty, that is, without merchantable size trees) selected from New York State forest lands and (ii) the total height and biomass of each individual tree was generated by Monte Carlo techniques that took into account the effect of diameter, merchantable height, species, site quality, geographical region, individual plot and finally, the random variation of tree height or biomass about its own (conditional) expected value.

To apply stratified sampling, one must use strata that are simple to define and sample, and sufficiently different from each other to make the sampling efficient. For simplicity, the sample plots were classified into three strata by the geographical region in which they happened to fall. Unfortunately, this resulted in strata

that were very similar to each other. Preliminary samples with plots selected by an allocation that was far from proportional, generated unstratified means (which ignored the plot stratification) that were not significantly different from the true mean (biomass per acre) of the population. This stratification obviously was not appropriate for the objectives of our study. Consequently, changes were required in the population tree biomass to make the differences between strata sufficiently large. This was accomplished by increasing (or decreasing) the biomass of each individual tree from a given stratum by an amount calculated by a formula of the form $(a+bd^2)$, where the constants a and b were stratum specific and d was the tree diameter.

More specifically, we have defined (i) the new Population 2 by using the values (a,b) equal to $(238, -2.46)$, $(60, .60)$ and $(-11, -1.34)$ for the trees of stratum 1, 2, and 3 respectively, (ii) the new Population 3 by using $(200, .40)$ for stratum 1 and $(-11, -1.34)$ for stratum 3, with the biomass of the trees of stratum 2 remaining unchanged and (iii) the new Population 4 where the values (a,b) were those equal to $(150, .30)$, $(10, -.50)$ and $(200, 1.50)$ for stratum 1, 2 and 3 respectively. As it will be later seen, the three strata of the new three populations become sufficiently different.

Stratified Two-Stage Sampling Procedure

The first stage of the sampling procedure consists of m_1 , m_2 and m_3 non-empty sample plots (clusters of trees) selected by simple random sampling from the 233, 188 and 246 non-empty plots of stratum 1, 2, and 3 respectively. There are also 64, 119 and 77 empty plots (without merchantable size trees) in these strata, but there was no point in selecting such plots in the sample. They were considered, however, when the estimate of the mean biomass per acre was calculated. In the second stage, a fixed percentage p of trees from each plot of the first stage sample was selected by simple random sampling without replacement. Because the multiplication of the percentage p with the number n_i of trees in the i -th plot is not necessarily an integer, a random number R from .00 to .99 was used to decide whether an additional sample tree is to be selected from the plot (when $R < \text{fractional part of the product } n_i p$) or no additional tree is to be selected otherwise.

For the purpose of the present study forty sets of 100 independent samples were selected by computer from each of the three strata. These forty sets represent forty different combinations of (i) first stage number m of sample plots ($m = 2, 5, 10, 15, 20, 30$) and (ii) second stage percentage p of sample trees ($p = 5, 10, 15, 20, 40, 60, 100$). Not all combinations of m and p values above were used, and the resulting sample size of the combinations used varied from 7 to 1020 trees. The specific combinations of m and p that were used are listed in Arabatzis (1986); they are not given here.

A stratified sample is now defined as a combination of three samples, one from each stratum. As there are 4000 samples from each stratum (40 sets of 100 simulated samples) there are theoretically a total of $4000^3 = 64,000,000,000$ possible stratified samples. To reduce the number of these samples to a manageable size, we have decided to consider stratified samples that (i) use the same percentage p in all strata, (ii) are generated by the same simulation run (that is, they have the same simulation run number) and (iii) consist of arbitrarily selected values m_1 , m_2 and m_3 . The specific types of stratified samples used will be given later.

Estimation Procedures for the Biomass Regressions

The various procedures to estimate the tree biomass regression functions are defined as combinations of (i) assumed form of the regression function, (ii) type of least squares method used and (iii) way by which the stratification of sample tree data is taken into account.

Two basic forms of regression function were used,

$$\text{form 1: } \hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2, \text{ and}$$

$$\text{form 2: } \hat{y} = \beta_1 + \beta_2 d^2 + \beta_3 d^2 h$$

where y = tree biomass and, as defined before, d = tree diameter and h = height. The selection of these forms was based on (i) the conclusions reached by Cunia and Gillespie (1985), Gillespie and Cunia (1986) and Michelakackis and Cunia (1985, 1986) and (ii) the fact that a term of the form bd^2 was added to each tree biomass of Populations 2, 3 and 4.

Four least squares methods were used, the ordinary least squares (OLS), ordinary weighted least squares (OWLS), modified least squares (MLS) and modified weighted least squares (MWLS). These methods are described in more detail in Cunia and Gillespie (1985), Gillespie and Cunia (1986), Michelakackis and Cunia (1985, 1986), Cunia (1986) and Arabatzis (1986). This detailed description is not repeated here. It suffices to say that (i) OLS is the usual least squares regression procedure applied to individual tree data without reference to the plot the tree was selected from, (ii) OWLS is the usual weighted least squares method where the conditional variance of y given d is assumed to be proportional to d^4 and the conditional variance of y given d and h is assumed to be proportional to $d^4 h^2$ and (iii) MLS and MWLS are the usual least and weighted least squares methods modified so as to take the cluster (plot) effect into account; these modifications consisting of the application of the weighted least squares method to the plot (not tree) data with the assumptions about the weights to use varying from MLS to MWLS.

We have defined nine procedures to take (or not to take) into account the effect of stratification, when the data of the sample trees are being used. These procedures are described in

detail by Arabatzis (1986) and for more information, the reader is referred to his work. However, because they are essential to the interpretation of the results obtained and conclusions reached from the analysis of these results, the nine procedures are summarized below.

Procedure 1 - The trees are assumed to have been selected by unstratified sampling and, thus, one common regression function is estimated, and then applied, to all three strata. This is the usual procedure used by foresters when one regression function is desired for all strata and stratified sampling is used to select the sample trees in order to insure a better representation of the trees from each stratum.

Procedure 2 - Stratified sampling is assumed, with the trees from one stratum selected independently from those of another stratum. Three biomass regression functions, one for each stratum, are estimated independently of each other and each regression is applied to its own stratum when the estimate of the mean biomass per acre is calculated. This is the right procedure to use when the sample trees were selected by the classical method of stratified random sampling.

Procedure 3 - Same as Procedure 2 but the three biomass regressions are calculated simultaneously by weighted least squares regression techniques that use dummy variables to (i) test null hypotheses about similarities between the regression functions of various strata and (ii) calculate regression coefficients that are common to several regressions, when these regression coefficients are not significantly different from each other. The least squares with dummy variable techniques are fully described by Cunia (1973) and the interested reader should refer to him for more details.

Procedure 4 - The three stratum biomass regression functions of Procedure 2 are first calculated as usual, and then, a single biomass regression function, applicable to all strata is defined as the weighted regression

$$\hat{Y} = w_1 r_1 + w_2 r_2 + w_3 r_3$$

where r_i is the regression of stratum i and the weight w_i is defined as the ratio of the total area of stratum i to the total area of all strata, for $i = 1, 2, 3$.

Procedure 5 - Similar to Procedure 4 but the weights w_i are defined as the ratio of the total number of trees in stratum i to the total number of trees in all strata, $i = 1, 2, 3$.

Procedure 6 - Similar to Procedure 4 but the three biomass regression functions that were averaged were calculated by Procedure 3 and not Procedure 2.

Procedure 7 - Similar to Procedure 6 but the weights w_i are those of Procedure 5 (based on number of trees) and not those of Procedure 4 (based on stratum areas).

Procedure 8 - A single biomass regression function is calculated for all three strata by the weighted least squares method, where the weights used are defined as functions of (i) area of stratum, (ii) sample size within a stratum and (iii) the usual tree values of d or d and h . More specifically, the weighted least squares method was applied as if, for each individual tree, the conditional variance of the biomass is (i) proportional to stratum area, (ii) inversely proportional to sample size within stratum and (iii) proportional to d^4 or $d^4 h^2$ as the case may be.

Procedure 9 - Similar to Procedure 8, but the weight of each individual tree is made proportional to stratum number of trees rather than stratum area.

Combining the form of the regression function (2) with the least squares method (4) and the procedure to take the tree stratification into account (9) yields a total of 72 procedures for the estimation of the biomass regression functions. This means that, for each simulated sample, we have calculated a total of 72 biomass regressions. To simplify the discussion we shall identify the three criteria of the estimation procedure as (i) form of regression function (ii) least squares estimation approach and (iii) procedure number (to take the tree stratification into account).

Note that, from among the nine procedures we have described, the second procedure is one which is theoretically unbiased, provided of course that the estimators of the means μ_i within stratum i are themselves unbiased. Consequently procedure 2 may be taken as the one to be compared with all of the remaining eight procedures.

Estimation Procedure for the Mean Biomass per Acre

To calculate z , the estimate of the overall mean biomass per acre of the entire population, we shall use the stratified sampling formula

$$z = Q_1 z_1 + Q_2 z_2 + Q_3 z_3$$

where, for stratum $i = 1, 2, 3$,

A_i = area of stratum i (acres)

$A = A_1 + A_2 + A_3$ = total area (in acres)

= total area (in acres) of the 927 one-fifth acre plots of our population

$Q_i = A_i/A$ = relative size of stratum i , and

z_i = estimate of the mean biomass per acre μ_i of stratum i .

Let us write now the biomass regression function of stratum i as

$$\hat{Y} = b_{i1} x_1 + b_{i2} x_2 + b_{i3} x_3$$

where

$x_1 = 1$ for all trees

$x_2 = d$ or d^2 for regression function form 1 or 2 respectively, and

$x_3 = d^2$ or d^2h for regression function form 1 or 2 respectively.

If the totals of the variables y , x_1 , x_2 and x_3 for stratum i are written as

τ_i = total of tree biomass y in stratum i

τ_{i1} = total of values x_1 in stratum i

= total number of trees in stratum i

τ_{i2} = total of values x_2 in stratum i

= sum of tree diameters d or squared tree diameters d^2 for regression function form 1 or 2 respectively, and

τ_{i3} = total of values x_3 in stratum i

= sum of squared tree diameters d^2 or crossproducts d^2h for regression function form 1 or 2 respectively,

then, the corresponding averages on a per acre basis are

τ_i/A_i = mean biomass per acre in stratum i

τ_{i1}/A_i = mean number of trees per acre in stratum i

τ_{i2}/A_i = mean sum of d or d^2 per acre as the case may be and

τ_{i3}/A_i = mean sum of d^2 or d^2h per acre as the case may be

Using the above notation it is easy to show that

(1) the estimate z_i of μ_i can be written as

$$z_i = b_{i1}(\tau_{i1}/A_i) + b_{i2}(\tau_{i2}/A_i) + b_{i3}(\tau_{i3}/A_i)$$

and

(2) the estimate z of the overall mean μ is

$$\begin{aligned} z &= Q_1 z_1 + Q_2 z_2 + Q_3 z_3 \\ &= (1/A) (b_{11}\tau_{11} + b_{12}\tau_{12} + b_{13}\tau_{13} + b_{21}\tau_{21} \\ &\quad + \dots + b_{33}\tau_{33}) \\ &= [B]' [\mu_x] \end{aligned}$$

where $[]$ and $[]'$ denotes matrices and transpose matrices

$$[B]' = [b_{11} \ b_{12} \ b_{13} \ b_{21} \ b_{22} \ b_{23} \ b_{31} \ b_{32} \ b_{33}]$$

and

$$[\mu_x]' = [(\tau_{11}/A) \ (\tau_{12}/A) \ \dots \ (\tau_{33}/A)]$$

Because $[\mu_x]$ is known without error, and if the covariance matrix of $[B]$ is estimated by $[S_{BB}]$, then, the variance of z is estimated by the expression

$$V = [\mu_x]' [S_{BB}] [\mu_x]$$

If the basic assumptions of the sampling and estimation procedures are strictly satisfied, z and V are unbiased estimators. As these assumptions are never satisfied, both z and V are biased. The objectives of this study is to find, among other things, to what extent z and V are biased.

Analysis Procedure

Each simulated sample may be viewed as a random experiment. For a given sampling procedure, sample size and allocation, the 100 simulated samples may also be viewed as 100 random outcomes, the result of 100, identically defined and statistically independent random experiments. Each random experiment generates 72 sets of random variables, one set for each estimation procedure. Each set contains (i) two random variables z and V (as defined in the previous section) that are taken as estimators of the mean biomass per acre μ and the variance of z (as an estimator of μ) respectively, (ii) two random intervals ($z \pm t\sqrt{V}$), with $t = 2.0$ and $t = 2.6$ and (iii) two multinomial random variables that take the values -1, 0 and 1, depending on whether the μ falls below, within or above the two random intervals above.

Because, for each estimation procedure and sampling method, sample size and allocation we have 100, identically distributed and statistically independent random variables z , we can use their average $\bar{z}$ and variance S_{zz} as estimators of the mean μ_z and variance σ_{zz} . As z is taken as an estimator of μ , (i) the sample statistic $(\bar{z} - \mu)$ may be used as an estimator of the bias of z , (ii) the sample statistic $S_{zz} = S_{zz}/100$ is an estimator of the variance of the bias $(\bar{z} - \mu)$ of z (as well as estimator of the variance of $\bar{z}$) and (iii) $t = (\bar{z} - \mu)/\sqrt{S_{zz}}$ can be used as a statistic to test the null hypothesis that the bias of z is equal to zero.

The variance of z is also estimated by the 100 sample values V as well as their average $\bar{V}$. This last value would normally be an unbiased estimator of the true variance of z whenever all the assumptions of the statistical model used in the corresponding sampling and estimation procedure are strictly satisfied. Another estimator of the same true variance of z , this time unbiased, is the statistic S_{zz} defined above. Consequently, the ratio V/S_{zz} , (or its square root), may be taken as an estimator of the relative amount by which V underestimates or overestimates the variance of z . Another measure of the goodness of V as an estimator of the variance of z is provided by the number of times (out of the 100 times) that the value μ fell below, within or above the 95 and 99 percent confidence intervals.

All of these values for all stratified sample sizes and estimation procedures considered in this study are listed in a series of tables by Arabatzis (1986). As an example, we have shown in Tables 1, 2, 3, and 4, for populations 1, 2, 3 and 4 respectively and for the 72 estimation procedures, the statistics $\bar{z}$, $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, $\sqrt{V}$, $\sqrt{V/S_{zz}}$, t and the number of times μ fell below, within and above the 95 and 99 percent confidence intervals, for the stratified sampling method with $m_1=30$, $m_2=20$, $m_3=10$ and $p=15$ percent.

The procedure to analyze the tables is to a large extent subjective. The conclusions were drawn by an ocular analysis of the values of $\bar{z}$, $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, $\sqrt{V/S_{zz}}$, etc. as listed in the table or as plotted in a set of graphs.

Analysis of Results - Part 1

To get an overall view of the type of results obtained, we have analyzed first, in relative detail, one specific stratified sample size and allocation, the case of $m_1=30$, $m_2=20$ and $m_3=10$ clusters selected at random from stratum 1, 2 and 3 respectively with 15 percent of their trees subsampled at random. The statistics $\bar{z}$, $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, $\sqrt{V}$, $\sqrt{V/S_{zz}}$, t and the number of times the confidence intervals fell above, over and below μ are listed by estimation procedure in Tables 1, 2, 3 and 4 for Populations 1, 2, 3 and 4 respectively. We shall successively analyze the bias of z , the precision and estimated precision of z and finally, the confidence intervals (that include the combined effect of the bias and estimated precision).

Analysis of the Bias of z

An ocular analysis of the bias $(\bar{z}-\mu)$ of Population 1 shows that, with the exception of the ordinary weighted least squares estimates by the regression form 2, the bias is not significantly different from zero; the values t are small, negative or positive values. And even when the bias is significant, the significance level is ordinarily above one percent. This is to be expected since the difference between the various strata is very small and any estimation procedure that gives too much weight to any given stratum could not unduly affect the bias. This seems to imply that, whenever the strata are sufficiently similar, the bias of the estimator z of μ would normally be close to, if not equal to zero.

The conclusions change, however, with the other three populations, where the strata were made, on purpose, to be different from each other. Starting with Procedure 1 (of a single unstratified biomass regression) the reader can verify that the sample bias is high in value and highly significant. Depending on the population considered, the bias is about 3, 7 and -5 percent of μ and the value of the test statistic t is about 10, 20 and -17. Consequently, the evidence strongly suggests that treating a stratified as an unstratified sample (when calculating the

biomass regression) is a poor procedure as it may seriously bias the estimates. This is to be expected on purely intuitive grounds.

With the Procedure 2 estimates, the results are different; the bias is small and not significantly different from zero. The one exception is with the estimates by the ordinary weighted least squares applied to the regression form 2. This is hard to explain. It may be something inherent to the way by which the estimates are calculated by the weighted least squares. It may be due to some specific (and unidentified) characteristic of our population. We do not know. But because (i) the weighted least squares estimates of the regression form 1 are statistically unbiased (ii) we expected statistically significant biased results to occur once in a while even where the procedure generating them is unbiased, and (iii) the value of the t -statistic is relatively low, about -2.34, we conclude that, in general Procedure 2 yields, as intuitively expected, unbiased results. The procedure estimates the mean μ_i of each stratum i separately and independently of the estimation of the mean μ_j of any other stratum j and then combines the three estimates z_1 , z_2 and z_3 into a single estimate of the overall mean μ by formulae which, from a theoretical point of view at least are unbiased.

Procedure 3 forces the regressions of the individual strata (of Procedure 2) to be parallel or identical for some simulated samples, when it is indeed known (in our case) that these regressions are not parallel or identical. This is a possible source of bias. As shown by an ocular analysis of Tables 2, 3 and 4, the results are mixed; the bias is not significantly different from zero for Population 2 but the bias is sometimes significant and sometimes not significant for the other two populations. This seems to imply that the third procedure may introduce a bias in the estimates, even though this bias is, most of the time relatively small, of the order of less than one percent of μ .

Procedures 4, 6 and 8, where a single regression function (common to all strata) is calculated by procedures using weights proportional to stratum areas (as measures of stratum sizes) yield high bias values that are significantly different from zero. Sometimes, the bias value may happen to be small and not significant. But this seems to occur at random and one cannot find a specific form of regression function or a least squares approach for which the bias would be consistently small. Things are different, however, when the measure of the stratum size is based on the total number of trees it contains. Then, the results are much better. Procedures 5 and 9 produce consistently small bias values which only occasionally are significantly different from zero, while for Procedure 7, the bias is significantly different from zero in 7 out of 24 cases. This last result is not unexpected since the single biomass regression used by Procedure 7 is based on three separate regressions which applied separately to the corresponding strata yield, in several instances, biased results. All this seems to imply that (i) using

TABLE 1: The statistics $\bar{z}$, $(\bar{z}-\mu)$, $\sqrt{s_{zz}}$, $\sqrt{v}$, the ratio $\sqrt{v/s_{zz}}$, t and the number of time μ fell below, within or above the 95% and 99% confidence intervals, as calculated by procedures 1, 2, 3, 4, 5, 6, 7, 8 and 9 for population 1; Sampling method: 15 percent of trees selected from 30, 20 and 10 plots of stratum 1, 2 and 3 respectively.

estimation									number of times					
model/ procedure		$\bar{z}$	$(\bar{z}-\mu)$	$\sqrt{s_{zz}}$	$\sqrt{\bar{v}}$	$\sqrt{\bar{v}/s_{zz}}$	t		95%			99%		
									b	w	a	b	w	a
1	111	114.989	-0.56	3.569	2.551	0.71	-1.57		5	84	11	2	94	4
2	121	115.096	-0.45	3.433	2.508	0.73	-1.32		7	85	8	1	95	4
3	131	115.057	-0.49	3.720	3.608	0.97	-1.32		2	94	4	1	96	3
4	141	115.065	-0.48	3.573	3.887	1.09	-1.36		0	97	3	0	98	2
5	211	115.770	0.22	2.575	1.887	0.73	0.86	11	84	5	4	93	3	
6	221	114.933	-0.61	2.464	1.964	0.80	-2.50	2	87	11	2	94	4	
7	231	115.745	0.19	2.711	2.663	0.98	0.72	5	92	3	0	99	1	
8	241	115.699	0.14	2.401	2.885	1.20	0.62	1	98	1	0	100	0	
9	112	115.164	-0.38	4.047	2.734	0.68	-0.95	10	80	10	4	88	8	
10	122	115.597	0.04	3.846	2.816	0.73	0.12	7	88	5	2	94	4	
11	132	115.139	-0.41	5.108	4.371	0.86	-0.80	3	92	5	1	97	2	
12	142	115.272	-0.27	4.598	4.782	1.04	-0.60	1	97	2	0	99	1	
13	212	115.724	0.17	2.965	2.022	0.68	0.59	12	81	7	8	88	4	
14	222	114.890	-0.65	2.814	2.092	0.74	-2.34	5	87	8	2	93	5	
15	232	115.639	0.09	3.384	3.084	0.91	0.27	8	89	3	3	96	1	
16	242	115.674	0.12	2.972	3.338	1.12	0.42	3	95	2	0	100	0	
17	113	114.969	-0.58	3.873	2.756	0.71	-1.50	6	85	9	3	92	5	
18	123	115.349	-0.20	3.476	2.554	0.73	-0.58	7	87	6	1	95	4	
19	133	115.184	-0.36	4.193	3.892	0.93	-0.87	4	92	4	2	95	3	
20	143	115.338	-0.21	3.690	3.905	1.06	-0.57	1	96	3	0	98	2	
21	213	115.771	0.22	2.934	2.046	0.70	0.76	11	83	6	7	90	3	
22	223	114.870	-0.67	2.614	2.008	0.77	-2.60	6	84	10	2	94	4	
23	233	115.848	0.29	2.943	2.773	0.94	1.01	7	90	3	1	98	1	
24	243	115.698	0.14	2.505	2.869	1.15	0.59	1	98	1	0	100	0	
25	114	115.252	-0.29	3.934	2.693	0.68	-0.76	11	79	10	6	89	5	
26	124	115.684	0.13	3.725	2.764	0.74	0.36	8	87	5	3	94	3	
27	134	115.262	-0.28	4.945	4.300	0.87	-0.58	3	93	4	1	97	2	
28	144	115.380	-0.16	4.451	4.683	1.05	-0.38	1	97	2	0	99	1	
29	214	115.812	0.26	3.021	2.020	0.67	0.87	13	77	10	9	87	4	
30	224	114.856	-0.69	2.737	2.047	0.75	-2.53	5	87	8	2	93	5	
31	234	115.783	0.23	3.476	3.095	0.89	0.67	9	88	3	2	97	1	
32	244	115.648	0.09	2.970	3.279	1.10	0.33	2	96	2	0	100	0	
33	115	115.189	-0.36	4.056	2.736	0.67	-0.89	11	78	11	4	88	8	
34	125	115.590	0.04	3.862	2.813	0.73	0.11	8	87	5	2	94	4	
35	135	115.201	-0.34	5.151	4.378	0.85	-0.68	3	92	5	1	97	2	
36	145	115.305	-0.24	4.645	4.784	1.03	-0.53	1	97	2	0	99	1	
37	215	116.073	0.52	3.083	2.032	0.66	1.70	15	78	7	10	85	5	
38	225	115.022	-0.52	2.817	2.092	0.74	-1.87	5	88	7	2	94	4	
39	235	116.048	0.49	3.581	3.124	0.87	1.39	9	88	3	3	97	0	
40	245	115.833	0.28	3.070	3.367	1.10	0.92	3	95	2	0	100	0	
41	116	115.046	-0.50	3.765	2.721	0.72	-1.34	6	85	9	5	92	3	
42	126	115.397	-0.15	3.428	2.545	0.74	-0.45	7	87	6	1	95	4	
43	136	115.251	-0.29	4.133	3.849	0.93	-0.72	4	92	4	2	95	3	
44	146	115.373	-0.17	3.739	3.899	1.04	-0.47	2	95	3	0	98	2	
45	216	115.836	0.28	2.992	2.018	0.67	0.96	14	79	7	8	88	4	
46	226	114.844	-0.70	2.574	1.997	0.78	-2.74	5	85	10	2	94	4	
47	236	115.816	0.26	2.921	2.737	0.94	0.91	7	90	3	1	98	1	
48	246	115.643	0.09	2.467	2.865	1.16	0.38	1	98	1	0	100	0	
49	117	114.998	-0.55	3.884	2.757	0.71	-1.42	6	85	9	3	92	5	
50	127	115.345	-0.20	3.483	2.553	0.73	-0.59	7	86	7	1	95	4	
51	137	115.225	-0.32	4.217	3.895	0.92	-0.77	4	92	4	2	95	3	
52	147	115.332	-0.21	3.688	3.905	1.06	-0.59	1	96	3	0	98	2	
53	217	116.094	0.54	3.047	2.050	0.67	1.79	14	80	6	8	89	3	
54	227	114.939	-0.61	2.609	2.006	0.77	-2.34	6	84	10	2	94	4	
55	237	115.990	0.44	3.013	2.768	0.92	1.46	9	89	2	2	97	1	
56	247	115.742	0.19	2.501	2.867	1.15	0.77	1	98	1	0	100	0	
57	118	115.651	0.10	3.672	2.508	0.68	0.28	10	83	7	5	91	4	
58	128	115.885	0.33	3.602	2.439	0.68	0.93	13	83	4	5	91	4	
59	138	115.706	0.15	3.858	3.542	0.92	0.41	4	93	3	2	96	2	
60	148	115.725	0.17	3.688	3.753	1.02	0.48	2	95	3	0	98	2	
61	218	115.457	-0.09	2.873	1.883	0.66	-0.32	12	82	6	7	89	4	
62	228	114.683	-0.86	2.669	1.878	0.70	-3.24	3	86	11	2	91	7	
63	238	115.417	-0.13	2.931	2.652	0.90	-0.45	7	90	3	2	96	2	
64	248	115.490	-0.06	2.584	2.760	1.07	-0.23	3	94	3	0	99	1	
65	119	115.583	0.03	3.753	2.512	0.67	0.09	12	79	9	5	91	4	
66	129	115.783	0.23	3.724	2.446	0.66	0.63	14	81	5	5	91	4	
67	139	115.651	0.10	3.959	3.546	0.90	0.26	5	92	3	2	96	2	
68	149	115.646	0.09	3.811	3.768	0.99	0.25	1	96	3	0	98	2	
69	219	115.710	0.16	2.934	1.872	0.64	0.55	16	79	5	8	88	4	
70	229	114.843	-0.70	2.744	1.890	0.69	-2.57	6	86	8	3	92	5	
71	239	115.682	0.13	2.990	2.638	0.88	0.44	8	89	3	3	95	2	
72	249	115.665	0.11	2.649	2.778	1.05	0.44	4	93	3	0	99	1	

TABLE 2: The statistics $\bar{z}$, $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, $\sqrt{V}$, the ratio $\sqrt{V}/S_{zz}$, t and the number of times μ fell below, within or above the 95% and 99% confidence intervals, as calculated by procedures 1, 2, 3, 4, 5, 6, 7, 8, and 9 for population 2; Sampling method: 15 percent of trees selected from 30, 20 and 10 plots of stratum 1, 2 and 3 respectively.

estimation model/ procedure		$\bar{z}$	$(\bar{z}-\mu)$	$\sqrt{S_{zz}}$	$\sqrt{V}$	$\sqrt{V}/S_{zz}$	t	number			of times		
								95%			99%		
								b	w	a	b	w	a
1	111	117.688	3.83	3.748	2.722	0.73	10.23	36	62	2	23	77	0
2	121	117.859	4.00	3.659	2.876	0.79	10.95	36	63	1	20	80	0
3	131	117.707	3.85	3.861	3.965	1.03	9.98	18	82	0	6	94	0
4	141	117.783	3.92	3.857	4.425	1.15	10.19	12	88	0	2	98	0
5	211	118.553	4.69	2.538	2.057	0.81	18.52	64	36	0	41	59	0
6	221	116.401	2.54	2.842	2.643	0.93	8.96	17	82	1	4	96	0
7	231	118.477	4.62	2.722	3.029	1.11	16.99	36	64	0	15	85	0
8	241	117.571	3.71	2.805	3.708	1.32	13.25	10	90	0	1	99	0
9	112	113.468	-0.38	4.047	2.734	0.68	-0.95	10	80	10	4	88	8
10	122	113.901	0.04	3.846	2.816	0.73	0.12	7	88	5	2	94	4
11	132	113.444	-0.41	5.108	4.371	0.86	-0.80	3	92	5	1	97	2
12	142	113.577	-0.27	4.598	4.782	1.04	-0.60	1	97	2	0	99	1
13	212	114.029	0.17	2.965	2.022	0.68	0.59	12	81	7	8	88	4
14	222	113.195	-0.65	2.814	2.092	0.74	-2.34	5	87	8	2	93	5
15	232	113.944	0.09	3.384	3.084	0.91	0.27	8	89	3	3	96	1
16	242	113.978	0.12	2.972	3.338	1.12	0.42	3	95	2	0	100	0
17	113	113.430	-0.42	3.937	2.861	0.73	-1.08	6	85	9	3	92	5
18	123	113.872	0.01	3.612	2.728	0.76	0.05	9	86	5	1	95	3
19	133	114.338	0.48	4.798	4.030	0.84	1.01	9	86	5	2	96	2
20	143	114.168	0.31	4.270	4.222	0.99	0.74	4	94	2	0	99	1
21	213	114.090	0.23	2.998	2.062	0.69	0.79	12	82	6	7	90	3
22	223	113.180	-0.67	2.736	2.212	0.81	-2.46	5	88	7	2	95	2
23	233	114.197	0.34	3.335	2.969	0.89	1.03	7	90	3	2	96	2
24	243	113.779	-0.07	2.853	3.165	1.11	-0.26	1	98	1	0	100	0
25	114	114.527	0.67	3.934	2.693	0.68	1.71	11	82	7	6	90	4
26	124	114.959	1.10	3.725	2.764	0.74	2.97	12	84	4	5	93	2
27	134	114.536	0.68	4.945	4.300	0.87	1.38	4	93	3	2	96	2
28	144	114.655	0.80	4.451	4.683	1.05	1.80	2	96	2	0	99	1
29	214	115.067	1.23	3.021	2.020	0.67	4.08	21	75	4	12	85	3
30	224	114.131	0.27	2.737	2.047	0.75	1.01	10	84	6	4	94	2
31	234	115.058	1.20	3.476	3.095	0.89	3.46	12	87	1	5	95	0
32	244	114.923	1.06	2.970	3.279	1.10	3.60	5	94	1	1	99	0
33	115	113.273	-0.58	4.056	2.736	0.67	-1.43	8	81	11	4	88	8
34	125	113.674	-0.18	3.862	2.813	0.73	-0.47	7	87	6	2	93	5
35	135	113.284	-0.56	5.151	4.378	0.85	-1.11	3	91	6	1	97	2
36	145	113.389	-0.46	4.645	4.784	1.03	-1.00	1	96	3	0	99	1
37	215	114.157	0.30	3.083	2.032	0.66	0.98	12	80	8	8	87	5
38	225	113.106	-0.74	2.817	2.092	0.74	-2.65	5	88	7	2	94	4
39	235	114.132	0.27	3.581	3.124	0.87	0.78	7	90	3	2	96	2
40	245	113.917	0.06	3.070	3.367	1.10	0.21	3	95	2	0	100	0
41	116	114.486	0.63	3.828	2.813	0.73	1.65	8	87	5	5	92	3
42	126	114.882	1.02	3.525	2.691	0.76	2.92	12	85	3	4	94	2
43	136	115.308	1.45	4.556	3.974	0.87	3.19	9	88	3	5	93	2
44	146	114.996	1.14	4.124	4.183	1.01	2.77	6	92	2	0	99	1
45	216	115.141	1.28	3.034	2.031	0.67	4.24	21	75	4	14	83	3
46	226	114.098	0.24	2.663	2.170	0.81	0.92	6	90	4	3	95	2
47	236	115.141	1.28	3.368	2.926	0.87	3.82	15	83	2	4	94	2
48	246	114.435	0.58	2.808	3.119	1.11	2.07	2	97	1	0	100	0
49	117	113.236	-0.61	3.947	2.862	0.72	-1.57	6	85	9	2	92	6
50	127	113.689	-0.16	3.624	2.724	0.75	-0.46	7	86	7	1	95	4
51	137	114.235	0.38	4.845	4.032	0.83	0.79	9	86	5	1	97	2
52	147	114.064	0.21	4.297	4.220	0.98	0.49	3	94	3	0	99	1
53	217	114.217	0.36	3.103	2.065	0.67	1.17	13	80	7	8	89	3
54	227	113.050	-0.80	2.728	2.205	0.81	-2.95	5	88	7	2	96	2
55	237	114.269	0.41	3.485	2.976	0.85	1.19	7	89	4	3	95	2
56	247	113.714	-0.14	2.837	3.158	1.11	-0.49	1	97	2	0	100	0
57	118	115.172	1.31	3.694	2.691	0.73	3.57	18	79	3	4	94	2
58	128	115.485	1.63	3.635	2.921	0.80	4.49	13	84	3	5	93	2
59	138	115.257	1.40	3.822	3.998	1.05	3.67	4	94	2	1	98	1
60	148	115.438	1.58	3.787	4.488	1.19	4.18	0	98	2	0	100	0
61	218	114.985	1.13	2.830	2.082	0.74	4.00	16	80	4	10	88	2
62	228	112.828	-1.02	2.843	2.751	0.97	-3.61	1	96	3	0	97	3
63	238	114.981	1.12	2.913	3.157	1.08	3.87	6	92	2	2	98	0
64	248	114.070	0.21	2.870	3.660	1.34	0.75	0	99	1	0	100	0
65	119	113.958	0.10	3.746	2.684	0.72	0.28	5	91	4	2	95	3
66	129	114.219	0.36	3.717	2.947	0.79	0.98	8	89	3	1	96	3
67	139	114.094	0.24	3.910	3.973	1.02	0.62	2	96	2	1	97	2
68	149	114.241	0.38	3.872	4.494	1.16	1.00	0	98	2	0	98	2
69	219	114.103	0.24	2.861	2.064	0.72	0.87	13	83	4	5	92	3
70	229	111.685	-2.16	2.989	2.811	0.95	-7.30	0	89	11	0	97	3
71	239	114.147	0.29	2.956	3.134	1.06	0.99	3	95	2	1	97	2
72	249	112.988	-0.86	3.003	3.911	1.30	-2.88	0	99	1	0	100	0

TABLE 3: The statistics $\bar{z}$, $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, $\sqrt{\bar{v}}$, the ratio $\sqrt{\bar{v}/S_{zz}}$, t and the number of times μ fell below, within or above the 95% and 99% confidence intervals, as calculated by procedures 1, 2, 3, 4, 5, 6, 7, 8 and 9 for population 3; Sampling method : 15 percent of trees selected from 30, 20 and 10 plots of stratum 1, 2 and 3 respectively.

estimation model/ procedure		$\bar{z}$	$(\bar{z}-\mu)$	$\sqrt{S_{zz}}$	$\sqrt{\bar{v}}$	$\sqrt{\bar{v}/S_{zz}}$	t	number of times					
								95%			99%		
								b	w	a	b	w	a
1	111	126.759	7.12	3.508	2.758	0.79	20.31	68	32	0	52	48	0
2	121	126.903	7.26	3.474	3.298	0.95	20.92	56	44	0	33	67	0
3	131	126.790	7.15	3.738	4.303	1.15	19.14	36	64	0	20	80	0
4	141	126.865	7.23	3.740	5.182	1.39	19.33	20	80	0	8	92	0
5	211	127.414	7.78	2.667	2.252	0.84	29.18	88	12	0	76	24	0
6	221	125.201	5.56	3.187	3.441	1.08	17.47	35	65	0	17	83	0
7	231	127.218	7.58	2.909	3.825	1.31	26.07	49	51	0	27	73	0
8	241	125.966	6.33	3.387	5.153	1.52	16.70	14	86	0	2	98	0
9	112	119.249	-0.38	4.047	2.734	0.68	-0.95	10	80	10	4	88	8
10	122	119.681	0.04	3.846	2.816	0.73	0.12	7	88	5	2	94	4
11	132	119.224	-0.41	5.108	4.371	0.86	-0.80	3	92	5	1	97	2
12	142	119.357	-0.27	4.598	4.782	1.04	-0.60	1	97	2	0	99	1
13	212	119.809	0.17	2.965	2.022	0.68	0.59	12	81	7	8	88	4
14	222	118.975	-0.65	2.814	2.082	0.74	-2.34	5	87	8	2	93	5
15	232	119.724	0.09	3.384	3.084	0.91	0.27	8	89	3	3	96	1
16	242	119.759	0.12	2.972	3.338	1.12	0.42	3	95	2	0	100	0
17	113	119.263	-0.37	3.985	2.854	0.72	-0.93	8	83	9	3	92	5
18	123	120.777	1.14	3.736	2.610	0.70	3.06	17	79	4	8	90	2
19	133	119.637	0.00	4.587	4.242	0.92	0.01	6	88	6	3	94	3
20	143	120.819	1.18	3.886	4.120	1.06	3.05	0	98	2	0	99	1
21	213	119.830	0.19	2.939	2.062	0.70	0.67	11	83	6	6	91	3
22	223	119.633	0.00	3.028	2.111	0.70	0.00	9	86	5	3	95	2
23	233	119.906	0.27	3.186	2.959	0.93	0.85	4	93	3	1	98	1
24	243	120.994	1.36	2.581	3.016	1.17	5.27	5	95	0	0	100	0
25	114	118.781	-0.85	3.934	2.893	0.68	-2.17	7	82	11	4	90	6
26	124	119.213	-0.42	3.725	2.764	0.74	-1.13	6	88	6	2	93	5
27	134	118.790	-0.84	4.945	4.300	0.87	-1.71	3	93	4	1	97	2
28	144	118.909	-0.72	4.451	4.683	1.05	-1.63	1	95	4	0	99	1
29	214	119.341	-0.29	3.021	2.020	0.67	-0.97	11	77	12	7	89	4
30	224	118.385	-1.24	2.737	2.047	0.75	-4.56	4	86	10	0	94	6
31	234	119.312	-0.32	3.476	3.095	0.89	-0.93	4	93	3	0	99	1
32	244	119.177	-0.45	2.970	3.279	1.10	-1.54	2	96	2	0	100	0
33	115	119.319	-0.31	4.056	2.736	0.67	-0.78	11	79	10	4	88	8
34	125	119.719	0.08	3.862	2.813	0.73	0.22	8	87	5	3	93	4
35	135	119.330	-0.30	5.151	4.378	0.85	-0.59	3	92	5	1	97	2
36	145	119.435	-0.19	4.645	4.784	1.03	-0.43	1	97	2	0	99	1
37	215	120.203	0.56	3.083	2.032	0.66	1.84	15	79	6	10	85	5
38	225	119.152	-0.48	2.817	2.092	0.74	-1.71	5	88	7	3	93	4
39	235	120.178	0.54	3.581	3.124	0.87	1.52	9	89	2	3	97	0
40	245	119.963	0.32	3.070	3.367	1.10	1.07	3	95	2	0	100	0
41	116	118.796	-0.83	3.862	2.807	0.72	-2.16	5	85	10	4	92	4
42	126	120.185	0.55	3.632	2.592	0.71	1.52	10	84	6	4	93	3
43	136	119.200	-0.43	4.492	4.162	0.93	-0.97	5	90	5	3	94	3
44	146	120.236	0.60	3.841	4.088	1.06	1.57	0	98	2	0	99	1
45	216	119.320	-0.31	2.999	2.032	0.68	-1.05	9	81	10	6	90	4
46	226	118.968	-0.66	2.912	2.082	0.72	-2.29	6	86	8	2	94	4
47	236	119.199	-0.43	3.161	2.903	0.92	-1.38	2	94	4	1	97	2
48	246	120.213	0.57	2.497	2.988	1.20	2.32	1	98	1	0	100	0
49	117	119.332	-0.30	3.995	2.855	0.71	-0.76	8	83	9	4	91	5
50	127	120.787	1.15	3.740	2.609	0.70	3.08	17	79	4	8	90	2
51	137	119.710	0.07	4.610	4.245	0.92	0.16	6	88	6	3	94	3
52	147	120.817	1.18	3.892	4.121	1.06	3.04	0	98	2	0	99	1
53	217	120.186	0.55	3.061	2.067	0.68	1.80	14	80	6	8	89	3
54	227	119.735	0.10	2.996	2.106	0.70	0.34	9	86	5	3	95	2
55	237	120.103	0.46	3.252	2.953	0.91	1.44	6	92	2	2	97	1
56	247	121.054	1.42	2.557	3.014	1.18	5.55	5	95	0	0	100	0
57	118	119.261	-0.37	3.791	2.741	0.72	-0.99	8	85	7	4	93	3
58	128	119.466	-0.16	3.773	3.231	0.86	-0.45	5	91	4	0	98	2
59	138	119.493	-0.14	3.996	4.305	1.08	-0.35	3	94	3	1	98	1
60	148	119.539	-0.09	3.942	5.105	1.30	-0.24	0	98	2	0	99	1
61	218	119.012	-0.62	3.020	2.254	0.75	-2.06	7	87	6	2	94	4
62	228	116.748	-2.88	3.132	3.356	1.07	-8.22	0	92	8	0	96	4
63	238	119.082	-0.55	3.119	3.822	1.23	-1.77	2	96	2	0	99	1
64	248	117.801	-1.83	3.268	5.032	1.54	-5.61	0	100	0	0	100	0
65	119	119.802	0.16	3.873	2.764	0.71	0.43	11	83	6	4	93	3
66	129	119.964	0.33	3.892	3.290	0.85	0.85	7	90	3	1	97	2
67	139	120.059	0.42	4.103	4.365	1.06	1.04	3	94	3	1	98	1
68	149	120.085	0.45	4.077	5.203	1.28	1.11	0	98	2	0	99	1
69	219	119.850	0.21	3.103	2.273	0.73	0.70	13	82	5	5	93	2
70	229	117.313	-2.32	3.284	3.441	1.05	-7.07	0	93	7	0	97	3
71	239	118.905	0.27	3.236	3.893	1.20	0.84	2	96	2	1	98	1
72	249	118.358	-1.27	3.455	5.154	1.49	-3.69	0	100	0	0	100	0

TABLE 4: The statistics $\bar{z}$, $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, $\sqrt{v}$, the ratio $\sqrt{v}/S_{zz}$, t and the number of times μ fell below, within or above the 95% and 99% confidence intervals, as calculated by procedures 1, 2, 3, 4, 5, 6, 7, 8 and 9 for population 4; Sampling method: 15 percent of trees selected from 30, 20 and 10 plots of stratum 1, 2 and 3 respectively.

estimation model/ procedure		$\bar{z}$	$(\bar{z}-\mu)$	$\sqrt{S_{zz}}$	$\sqrt{\bar{v}}$	$\sqrt{\bar{v}/S_{zz}}$	t	number of times					
								95%			99%		
								b	w	a	b	w	a
1	111	131.543	-5.00	3.577	2.747	0.77	-16.78	0	46	54	0	58	42
2	121	131.730	-5.81	3.405	3.107	0.91	-17.08	0	53	47	0	74	26
3	131	131.850	-5.89	3.888	4.251	1.09	-15.17	1	75	24	0	91	9
4	141	131.739	-5.80	3.699	5.003	1.35	-15.70	0	83	17	0	94	6
5	211	132.291	-5.25	2.874	2.202	0.77	-18.29	0	33	67	0	54	46
6	221	131.374	-6.17	3.076	2.994	0.97	-20.07	0	46	54	0	69	31
7	231	132.231	-5.31	3.094	3.637	1.18	-17.18	0	72	28	0	90	10
8	241	132.169	-5.37	3.019	4.621	1.53	-17.81	0	87	13	0	96	4
9	112	137.162	-0.38	4.047	2.734	0.68	-0.95	10	80	10	4	88	8
10	122	137.594	0.04	3.846	2.816	0.73	0.12	7	88	5	2	94	4
11	132	137.137	-0.41	5.108	4.371	0.86	-0.80	3	92	5	1	97	2
12	142	137.270	-0.27	4.598	4.782	1.04	-0.60	1	97	2	0	99	1
13	212	137.722	0.17	2.965	2.022	0.68	0.59	12	81	7	8	88	4
14	222	136.888	-0.65	2.814	2.092	0.74	-2.34	5	87	8	2	93	5
15	232	137.636	0.09	3.384	3.084	0.91	0.27	8	89	3	3	96	1
16	242	137.671	0.12	2.972	3.338	1.12	0.42	3	95	2	0	100	0
17	113	137.125	-0.42	3.944	2.855	0.72	-1.07	6	85	9	3	92	5
18	123	136.604	-0.94	3.731	2.625	0.70	-2.53	6	80	14	1	94	5
19	133	137.135	-0.41	10.377	6.385	0.62	-0.40	6	86	8	3	94	3
20	143	136.740	-0.80	6.686	5.178	0.77	-1.21	1	95	4	0	99	1
21	213	137.763	0.21	2.979	2.063	0.69	0.73	12	82	6	7	90	3
22	223	136.447	-1.10	2.828	2.132	0.75	-3.89	5	86	9	1	94	5
23	233	137.805	0.25	3.338	2.976	0.89	0.77	7	90	3	2	97	1
24	243	136.858	-0.68	2.651	3.021	1.14	-2.60	0	98	2	0	99	1
25	114	135.279	-2.26	3.934	2.693	0.68	-5.76	5	73	22	1	88	11
26	124	135.711	-1.83	3.725	2.764	0.74	-4.93	3	84	13	0	95	5
27	134	135.288	-2.25	4.945	4.300	0.87	-4.57	3	87	10	0	97	3
28	144	135.407	-2.13	4.451	4.683	1.05	-4.81	0	94	6	0	98	2
29	214	135.839	-1.70	3.021	2.020	0.67	-5.65	2	78	20	0	88	12
30	224	134.883	-2.66	2.737	2.047	0.75	-9.73	0	70	30	0	85	15
31	234	135.810	-1.73	3.476	3.095	0.89	-5.00	0	91	9	0	97	3
32	244	135.675	-1.87	2.970	3.279	1.10	-6.30	0	94	6	0	98	2
33	115	137.230	-0.31	4.056	2.736	0.67	-0.78	11	79	10	4	88	8
34	125	137.630	0.08	3.862	2.813	0.73	0.22	8	87	5	3	93	4
35	135	137.241	-0.30	5.151	4.378	0.85	-0.59	3	92	5	1	97	2
36	145	137.346	-0.20	4.645	4.784	1.03	-0.43	1	97	2	0	99	1
37	215	138.114	0.56	3.083	2.032	0.66	1.84	15	79	6	10	85	5
38	225	137.063	-0.48	2.817	2.092	0.74	-1.72	5	88	7	3	93	4
39	235	138.089	0.54	3.581	3.124	0.87	1.51	9	88	3	3	97	0
40	245	137.874	0.32	3.070	3.367	1.10	1.06	3	95	2	0	100	0
41	116	135.248	-2.29	3.837	2.808	0.73	-5.99	4	77	19	0	90	10
42	126	134.907	-2.63	3.614	2.606	0.72	-7.30	1	78	21	0	86	14
43	136	135.315	-2.23	9.781	6.094	0.62	-2.28	3	83	14	1	93	6
44	146	135.041	-2.50	6.396	5.039	0.79	-3.92	1	89	10	0	97	3
45	216	135.842	-1.70	3.027	2.033	0.67	-5.63	2	76	22	1	88	11
46	226	134.514	-3.03	2.699	2.101	0.78	-11.24	0	68	32	0	87	13
47	236	135.726	-1.82	3.390	2.931	0.86	-5.37	1	90	9	0	97	3
48	246	134.963	-2.58	2.572	2.996	1.17	-10.05	0	89	11	0	98	2
49	117	137.190	-0.35	3.953	2.856	0.72	-0.90	6	85	9	4	91	5
50	127	136.609	-0.93	3.738	2.623	0.70	-2.51	6	80	14	1	94	5
51	137	137.197	-0.35	10.386	6.381	0.61	-0.34	6	86	8	3	94	3
52	147	136.740	-0.80	6.651	5.176	0.78	-1.21	1	95	4	0	99	1
53	217	138.119	0.57	3.084	2.068	0.67	1.86	14	80	6	9	88	3
54	227	136.549	-0.99	2.845	2.126	0.75	-3.51	5	86	9	1	94	5
55	237	138.021	0.47	3.479	2.982	0.86	1.36	10	88	2	4	95	1
56	247	136.928	-0.61	2.680	3.025	1.13	-2.31	0	98	2	0	99	1
57	118	135.568	-1.97	3.819	2.793	0.73	-5.18	2	79	19	1	88	11
58	128	135.929	-1.61	3.744	3.236	0.86	-4.32	1	90	9	0	96	4
59	138	135.560	-1.98	4.217	4.453	1.06	-4.71	1	91	8	1	96	3
60	148	135.670	-1.87	4.105	5.283	1.29	-4.57	1	95	4	0	99	1
61	218	135.404	-2.14	3.036	2.291	0.75	-7.06	2	74	24	1	89	10
62	228	135.056	-2.49	3.425	3.132	0.91	-7.27	0	88	12	0	95	5
63	238	135.301	-2.24	3.201	3.871	1.21	-7.02	0	94	6	0	98	2
64	248	135.810	-1.73	3.165	4.844	1.53	-5.49	0	98	2	0	100	0
65	119	137.477	-0.07	3.893	2.778	0.71	-0.18	9	81	10	2	92	6
66	129	137.820	0.27	3.864	3.189	0.83	0.71	7	89	4	1	97	2
67	139	137.426	-0.12	4.274	4.401	1.03	-0.28	2	93	5	1	97	2
68	149	137.519	-0.02	4.186	5.207	1.24	-0.07	1	97	2	0	99	1
69	219	137.620	0.07	3.091	2.252	0.73	0.24	10	84	6	3	93	4
70	229	137.267	-0.28	3.431	3.052	0.89	-0.82	2	94	4	0	97	3
71	239	137.452	-0.09	3.249	3.780	1.16	-0.29	1	97	2	0	99	1
72	249	138.005	0.45	3.158	4.721	1.49	1.45	0	100	0	0	100	0

Table 5: Precision $\sqrt{S_{zz}}$ of estimates z by (i) form of regression function, (ii) least squares estimation approach (iii) procedure for taking into account the stratum effect and (iv) population.

==POPULATION 1==											
Least squares method	Regression form	Procedure number									
		1	2	3	4	5	6	7	8	9	
OLS	1	3.57	4.05	3.87	3.93	4.06	3.76	3.88	3.67	3.75	
OWLS	1	3.43	3.85	3.48	3.73	3.86	3.43	3.48	3.60	3.72	
MLS	1	3.72	5.11	4.19	4.94	5.15	4.13	4.22	3.86	3.96	
MWLS	1	3.57	4.60	3.69	4.45	4.65	3.74	3.69	3.69	3.81	
OLS	2	2.57	2.96	2.93	3.02	3.08	2.99	3.05	2.87	2.93	
OWLS	2	2.46	2.81	2.61	2.74	2.82	2.57	2.61	2.67	2.74	
MLS	2	2.71	3.38	2.94	3.48	3.58	2.92	3.01	2.93	2.99	
MWLS	2	2.40	2.97	2.50	2.97	3.07	2.47	2.50	2.58	2.65	
==POPULATION 2==											
OLS	1	3.75	4.05	3.94	3.93	4.06	3.83	3.95	3.69	3.75	
OWLS	1	3.66	3.85	3.61	3.73	3.86	3.53	3.62	3.64	3.72	
MLS	1	3.86	5.11	4.80	4.94	5.15	4.56	4.84	3.82	3.91	
MWLS	1	3.86	4.60	4.27	4.45	4.65	4.12	4.30	3.79	3.87	
OLS	2	2.54	2.96	3.00	3.02	3.08	3.03	3.10	2.83	2.86	
OWLS	2	2.84	2.81	2.74	2.74	2.82	2.66	2.73	2.84	2.97	
MLS	2	2.72	3.38	3.34	3.48	3.58	3.37	3.49	2.91	2.96	
MWLS	2	2.80	2.97	2.85	2.97	3.07	2.81	2.84	2.87	3.00	
==POPULATION 3==											
OLS	1	3.51	4.05	3.99	2.94	4.06	3.88	4.00	3.79	3.87	
OWLS	1	3.47	3.85	3.74	3.03	3.86	3.63	3.74	3.77	3.89	
MLS	1	3.74	5.11	4.59	3.19	5.15	4.49	4.61	4.00	4.10	
MWLS	1	3.74	4.60	3.89	2.58	4.65	3.84	3.89	3.94	4.08	
OLS	2	2.67	2.96	2.94	3.93	3.08	3.00	3.06	3.02	3.10	
OWLS	2	3.19	2.81	3.03	3.73	2.81	2.91	3.00	3.13	3.28	
MLS	2	2.91	3.38	3.19	4.94	3.58	3.16	3.25	3.12	3.24	
MWLS	2	3.39	2.97	2.58	4.45	3.07	2.50	2.56	3.27	3.46	
==POPULATION 4==											
OLS	1	3.58	4.05	3.94	3.93	4.06	3.84	3.95	3.82	3.89	
OWLS	1	3.40	3.85	3.73	3.73	3.86	3.61	3.74	3.74	3.86	
MLS	1	3.89	5.11	10.38	4.94	5.15	9.78	10.39	4.22	4.27	
MWLS	1	3.70	4.60	6.69	4.45	4.65	6.40	6.65	4.10	4.19	
OLS	2	2.87	2.96	2.98	3.02	3.08	3.03	3.08	3.04	3.09	
OWLS	2	3.08	2.81	2.83	2.74	2.82	2.70	2.84	3.42	3.43	
MLS	2	3.09	3.38	3.34	3.48	3.58	3.39	3.48	3.20	3.25	
MWLS	2	3.02	2.97	2.65	2.97	3.07	2.57	2.68	3.16	3.16	

weighted regressions with weights based on area as a measure of stratum size is not a good procedure and (ii) using weighted regression with weights based on number of trees as a measure of stratum size may lead, although not always, to unbiased results.

To summarize the above discussion it appears that (i) Procedure 2 is always unbiased, (ii) Procedure 3 is almost always unbiased and (iii) if single biomass regressions applicable to all strata are desired, the individual stratum data (Procedures 8 and 9) or regressions (Procedures 4, 5, 6 and 7) should be weighted by number of trees, not area of the strata.

Analysis of Precision $\sqrt{S_{zz}}$ and Accuracy of z

To analyze the precision of z we could have used Tables 1, 2, 3 and 4. The analysis is facilitated, however, when all the estimates $\sqrt{S_{zz}}$ of the precision of z are rearranged in a separate Table 5 by least squares estimation approach, regression function form, procedure number and population.

Because Procedure 2 is the standard against which the other procedures are usually compared, let us start with its analysis. An ocular scan of the results in Table 5 shows that, as expected, (i) the weighted least squares estimators are slightly more precise than the least squares ones, (ii) the modified estimators are less precise than the ordinary ones and (iii) the estimators based on diameter and height (regression form 2) are much more precise than the ones based on diameter alone (regression form 1). These conclusions seem to hold approximately true for the other procedures as well, although sometimes, the weighted least squares estimators turn out to be less precise, than the corresponding least squares.

A look at the precision of the estimators calculated by Procedure 1 (using a single biomass regression function for all strata) shows that it is higher (that is, the value of $\sqrt{S_{zz}}$ is lower) than that of the estimators calculated by Procedure 2 (using a different biomass regression for each stratum). This seems surprising until we realize that the estimates of the single regression function are based on a much larger number of sample trees and, thus are much more precise

than the estimates of the three individual regressions of Procedure 2. On the other hand the estimators of Procedure 1 are less accurate. If we measure the accuracy by the mean square error expression $(S_{zz} + (\bar{z} - \mu)^2)$, and because the bias by Procedure 2 is generally smaller, the mean square error of the estimators by Procedure 1 turns out to be higher in value, and, thus, the estimators turn out to have a lower accuracy.

The estimates based on Procedure 3 (where parallel or identical regressions are used whenever warranted by the results of the significance tests) are generally more precise than those based on Procedure 2. Occasionally, however, they are much less precise. For example, the precision of the MLS estimators using regression form 1 of population 4 is extremely low. This inconsistency is hard to explain. Furthermore, the bias of the estimators by Procedure 3 is, in some instances, relatively large and significantly different from zero. Consequently, the estimators calculated by Procedure 2 seem to be more consistent and, on the average more accurate than those calculated by Procedure 3.

The precision of the estimators derived by Procedures 4 and 5 is very close to that of Procedure 2, and the precision of the estimators derived by Procedures 6 and 7 is very close to that of Procedure 3. This seems to imply that it does not matter whether we use separate regressions, one for each stratum, or a single regression for all strata, the weighted average of these separate regressions. However, because (i) there is always a real possibility of introducing a large bias when using Procedures 4 and 6 and (ii) it is questionable whether good estimators of the number of trees by strata (used as weights) can ever be obtained, Procedures 2 and 3 are to be preferred to Procedures 4, 5, 6 and 7.

Finally, the estimates based on Procedures 8 and 9 seem to have relatively good precision for population 1. For the other three populations the results seem to be mixed. As the estimates based on Procedure 8 are biased and the ones based on Procedure 9 require the knowledge of the total number of trees contained in each stratum, there seems to be no advantage in using them in preference to those generated by Procedure 2.

Analysis of the Estimated Precision $\sqrt{V}$ of z

The sample based estimator of the precision of z is the sample variance V , or its square root, the standard error $\sqrt{V}$. Because S_{zz} is an unbiased estimator of the variance of z (which is independent of the assumptions of the estimation model), the ratio V/S_{zz} or its square root $\sqrt{V/S_{zz}}$ can be used as a measure of the goodness of the estimator V . We shall use the square root. And to facilitate the analysis, the values of the ratios $\sqrt{V/S_{zz}}$ of Tables 1, 2, 3 and 4 were rearranged in Table 6 by population, regression function form, procedure number and least squares estimation approach.

As the reader can verify by an ocular analy-

sis of the results listed in Table 6, the ratio $\sqrt{V/S_{zz}}$ is consistently much lower in value than 1 for all ordinary least and weighted least squares estimators, with slightly lower values for the least than weighted least squares. This shows that V is a gross underestimate of the variance of z . On the average, the value of the ratio is about .70 for the OLS and about .78 for the OWLS estimators of all procedures. The statistic V of the modified least squares is also an underestimate of the variance of z ; on the average, the value of the ratio is about .90. On the other hand, the ratio for the modified weighted least squares estimators is over 1; this means that the variance of the MWLS estimators is somewhat overestimated by V .

Analysis of the Confidence Intervals of μ

The confidence intervals are a good expression of the combined effect of precision and bias; and for the layman, a much more intuitive and meaningful measure of the validity and accuracy of the estimators. A look at the number of times the 95 and 99 percent confidence intervals fall above, over or below μ leads to the conclusion that the confidence intervals are (i) badly off when the estimation procedures are based on the ordinary least and weighted least squares and (ii) relatively all right when the estimates are calculated by the modified procedures. It should be noted here that the proportion of times the confidence interval statements were found to be right is significantly different from the expected 95 or 99 times for the ordinary but not significantly different for the modified techniques.

Analysis of Results - Part 2

In the previous section we have made a detailed analysis of the results obtained when the stratified sample was of a specific size and allocation. The objective of the present section is to analyze other sample sizes and allocations and see whether the main calculations reached in the previous section apply in these cases as well; or, to see what changes, if any, should we make to our conclusions when the percentage p of trees selected from a sample cluster changes from the value $p=15$ used, and the number of clusters selected per stratum 1, 2, and 3 is different from that of 30, 20 and 10 respectively of the previous section.

To facilitate our discussion we shall define a sampling method as consisting of a specific percent p and a specific number of sample clusters m_1 , m_2 and m_3 . Because of the limitations due to computer time and costs, we have (i) selected 21 specific sampling methods (in addition to the one analyzed in the previous section) and (ii) eliminated the estimators by Procedures 3, 6, and 7 that required an extremely long computer time and did not seem to present any real advantage over Procedures 2, 4 and 5 respectively. The 22 sampling methods analyzed in this study are listed in Table 7. Note that the

Table 6 - The sample values of the ratio $\sqrt{V/S_{zz}}$ arranged by population, regression equation form, procedure number and least squares estimation approach.

Estimation Approach	Form of Regression	Population	Procedure Number								
			1	2	3	4	5	6	7	8	9
OLS	1	1	.71	.68	.71	.68	.67	.72	.71	.68	.67
		2	.73	.68	.73	.68	.67	.73	.72	.73	.72
		3	.79	.68	.72	.68	.67	.72	.71	.72	.71
		4	.77	.68	.72	.68	.67	.73	.72	.73	.71
	2	1	.73	.68	.70	.67	.66	.67	.67	.66	.64
		2	.81	.68	.69	.67	.66	.67	.67	.74	.72
		3	.84	.68	.70	.67	.66	.68	.68	.75	.73
		4	.77	.68	.69	.67	.66	.67	.67	.75	.73
OWLS	1	1	.73	.73	.73	.74	.73	.74	.73	.68	.66
		2	.79	.73	.76	.74	.73	.76	.75	.80	.79
		3	.95	.73	.70	.74	.73	.71	.70	.80	.85
		4	.91	.73	.70	.74	.73	.72	.70	.86	.83
	2	1	.80	.74	.74	.75	.74	.78	.77	.70	.69
		2	.93	.74	.81	.75	.74	.81	.81	.97	.95
		3	1.08	.74	.70	.75	.74	.72	.70	1.07	1.05
		4	.97	.74	.75	.75	.74	.78	.75	.91	.89
MLS	1	1	.97	.86	.93	.87	.85	.93	.92	.92	.90
		2	1.03	.86	.84	.87	.85	.83	.83	1.05	1.02
		3	1.15	.86	.92	.87	.85	.92	.92	1.08	1.06
		4	1.09	.86	.62	.87	.85	.61	.61	1.06	1.03
	2	1	.98	.91	.94	.89	.87	.94	.92	.90	.80
		2	1.11	.91	.89	.89	.87	.85	.85	1.08	1.06
		3	1.31	.91	.93	.89	.87	.91	.91	1.23	1.20
		4	1.18	.91	.89	.89	.87	.86	.86	1.21	1.16
MWLS	1	1	1.09	1.04	1.06	1.05	1.03	1.04	1.06	1.02	.99
		2	1.15	1.04	.99	1.05	1.03	1.01	.98	1.19	1.16
		3	1.39	1.04	1.06	1.05	1.03	1.06	1.06	1.30	1.28
		4	1.35	1.04	.77	1.05	1.03	.79	.78	1.29	1.24
	2	1	1.20	1.12	1.15	1.10	1.10	1.16	1.15	1.07	1.05
		2	1.32	1.12	1.11	1.10	1.10	1.11	1.11	1.34	1.30
		3	1.52	1.12	1.17	1.10	1.10	1.20	1.18	1.54	1.49
		4	1.53	1.12	1.14	1.10	1.10	1.17	1.13	1.53	1.49

sampling method 11 is the method considered in the previous section.

For each population and each sampling method we have calculated the same set of basic statistics as that shown in Table 1. To facilitate the analysis, however, the statistics from the various tables were arranged and rearranged several times in order to answer specific questions or verify specific conclusions reached in the analysis of the previous section. All these tables are listed in Arabatzis (1986).

To better see how the conclusions of the previous section change, we have summarized them all as a set of five main conclusions. Each conclusion will be stated separately and its erality analyzed in view of the new information provided by the results of the 21 additional sampling methods.

Conclusion 1 - (sampling method 11). The estimators by all procedures are unbiased for population 1, the only population for which the differences among strata are negligibly small. There is one exception, however, that of the OWLS estimators of the regression form 2; but their bias is thought to be due to sampling error. For the other three populations, only the procedures 2, 5, and 9 yield unbiased estimators; the procedure 2 that is expected to be unbiased on theoretical considerations, and the two procedure 5 (derived from procedure 2) and 9 that use weights based on stratum total number of trees. Procedures 3 and 7 yield mixed results; the least squares but not the weighted least squares estimators are generally unbiased. Procedure 1 (which ignores the effect of stratification) and procedures 4, 6 and 8 (that are all based on weights defined in terms of stratum areas) yield biased results.

Table 7 - The number of clusters m_1 , m_2 and m_3 of stratum 1, 2 and 3 respectively and the percent p of trees subsampled for the twenty-two stratified sampling methods considered in this study

Sampling Method					Sampling Method				
m_1	m_2	m_3	p		m_1	m_2	m_3	p	
1	5	5	5	15	12	30	30	30	15
2	5	10	20	15	13	5	5	5	5
3	5	15	30	15	14	15	15	15	5
4	10	20	30	15	15	30	20	10	5
5	15	15	15	15	16	30	30	30	5
6	15	20	5	15	17	30	15	5	5
7	15	30	5	15	18	5	5	5	60
8	20	10	5	15	19	15	15	15	60
9	20	30	10	15	20	30	20	10	60
10	30	15	5	15	21	30	30	30	60
11	30	20	10	15	22	30	15	5	60

The analysis of the bias of the entire set of 22 sampling methods shows similar results. We still have the estimators of population 1 unbiased, except for the OWLS estimators by regression form 2. For the other populations, procedures 2 and 5 yield unbiased estimators; but now, some of the results by the procedure 9 are biased. And as before, the bias of the estimators calculated by the procedures 1, 4 and 8 is significantly different from zero with many of the values t being very large.

Conclusion 2 - (sampling method 11). Within a given procedure (i) the weighted least squares estimators are, more often than not, more precise, but not by much, than the least squares ones, (ii) the ordinary least and weighted least squares approaches yield always more precise estimators than the modified ones and (iii) the estimators based on the regression of biomass on diameter and height are always more precise than those based on the regression of biomass on diameter alone.

The analysis of the results from the entire set of 22 sampling methods shows that, with the exception of two special cases that extend them, these conclusions are verified. The two exceptions refer to (i) the case of small number of sample clusters per stratum ($m=5$ results in modified least and weighted least squares regressions having only two degrees of freedom) where the estimators by the modified procedures are much less precise than those by the ordinary procedures and (ii) the case of small number of sample trees per cluster (when $p=5$, the average number of trees per cluster is less than 2) where the precision of the modified least squares method approaches that of the ordinary least squares. This seems to imply that the MLS and MWLS estimators should only be used when there are at least ten sample clusters per stratum.

Conclusion 3 - (sampling method 11). With the possible exception of population 1 (where the strata are very similar), it seems that, although

the estimators by procedure 2 are less precise than those by procedures 1 and 3, they are usually more accurate. Similarly, procedures 5 and 9 may yield slightly better estimators than procedures 2 but they require the knowledge of the total number of trees contained in each stratum. Finally, all of the remaining procedures may be more precise but, because their bias is generally large and, sometimes, they are inconsistent, their accuracy is low.

These conclusions are all verified by the analysis of the results from the additional 21 sampling methods.

Conclusion 4 - (sampling method 11). The sample-based statistic V grossly underestimates the variance of z when the ordinary least or weighted least squares techniques are used; the underestimation is much smaller with the modified least squares approach and for the modified weighted least squares, V overestimates the variance of z by a relatively small amount.

The analysis of the other 21 additional sampling methods show that, with some exceptions that may be viewed more as extensions, these conclusions hold true. When the number of sample trees per cluster is small (the case of $p=5$ which results in an average of less than 2 trees per cluster) the underestimation of the variance of z becomes smaller, and for the case of OWLS estimators it becomes negligibly small. When the number of sample trees per cluster increases to about 20 (the case of $p=60$) the underestimation of the precision of the OLS and OWLS estimators becomes much larger; often, the standard deviation of z is estimated to less than 50 percent of its value. As far as the modified least and weighted least squares methods are concerned, there seems to be no effect of p on the underestimation or overestimation of the precision.

Conclusion 5 - (sampling method 11). Taking into consideration the bias, precision and estimated precision, it seems that the best overall estimation procedure is procedure 2; it is unbiased, generally less precise but more accurate than procedures 5 and 9 which, although they may yield about equally good results, would require the additional knowledge of stratum size expressed as total number of trees. This conclusion holds true for the additional 21 sampling methods as well.

Summary Comments

We have described a specific stratified, two-stage sampling method of tree selection for construction of tree biomass tables for use with forest resource inventory. This method consists of (i) a first stage sample of m_1 , m_2 and m_3 clusters (plots) selected from stratum 1, 2 and 3 respectively of a given forest population and (ii) a second stage subsample of p percent of the trees contained in each sample cluster of the first stage. The selection procedure in both stages is simple random sampling without replacement. To facilitate the discussion, we have

defined a sampling method as consisting of a specific sample size and allocation, that is, a specific set of sampling parameters m_1 , m_2 , m_3 and p .

Using simulation techniques, these sampling methods were then repeatedly applied to a given forest population of 22,723 trees distributed in 927 one-fifth acre sample plots. These plots were part of the New York State forest inventory system carried out by the Northeastern Forest Experiment Station. The trees were measured in the field for their diameter and merchantable height, but not measured for their total height and biomass (green weight above ground); these were generated by a Monte Carlo process that preserved the effect (on the simulated value of the tree height and biomass) of such factors as species, diameter, merchantable height, site, geographical region, plot and inherent random variation. Although artificially constructed, this population of trees is expected to imitate, with sufficient accuracy, a natural forest; that is, is expected to have all of the basic characteristics of forest populations as found in nature. The forest population so constructed was then divided, somewhat arbitrarily into three geographical regions called strata. As the biomass regressions of the three resulting strata were quite similar, we have constructed three additional forest populations with biomass regression functions that were made to vary greatly from stratum to stratum.

The computer simulation process of sampling was repeated one hundred times and, thus, for each sampling method we have obtained one hundred different samples of trees. The tree biomass regression function of each sample was then estimated with the use of seventy-two statistical models, the combinations of four estimation approaches (OLS, OWLS, MLS and MWLS), two regression function forms (1 and 2) and nine procedures that took the stratum effect into account; procedure 1 that ignores the stratum effect and calculates one regression function for all strata, procedure 2 that calculates independent regressions for each stratum separately, procedure 3, similar to procedure 2, that uses dummy variables techniques to calculate individual stratum regressions that are not statistically independent of each other and the remaining six procedures that calculate single regressions for use in all strata by either (i) appropriately averaging the individual regressions of procedures 2 and 3 or (ii) appropriately weighing the information from sample tree data when the weighted least squares method is being applied.

The seventy-two regression functions of each sample of a representative set of sampling methods were then applied to the tree data of the four populations to calculate estimators z of (the average biomass per acre) and estimators S_{zz} and V of the variance of z . The estimator S_{zz} is calculated from the one hundred, identically distributed values z of the same sampling and estimation procedure, and the estimator V is calculated from the data of a single sample, under the assumptions of the statistical model

used. V is the only estimator of the variance of z that one can obtain in real life, but this estimator is generally biased as the assumptions of the model are rarely satisfied. On the other hand, S_{zz} is an unbiased estimator of the variance of z but it is seldom, if ever, known in real life. Of course, we can calculate S_{zz} by deliberately sampling the same population several times by the same sampling method, or we can deliberately split the sample data into several parts and use the variation between the estimates z of the parts as the basis for the calculation of S_{zz} .

Because the true value μ is known and we have one hundred, statistically independent and identically distributed sets of random variables z , V and $(z + t\sqrt{V})$, we are able to study the probability behavior of these variables for any given combination of sampling method and estimation procedure. In particular we can estimate the bias of z by the quantity $(z - \mu)$ and the accuracy of z by the mean square error defined as the sum of the variance S_{zz} and squared bias $(\bar{z} - \mu)^2$. We can also test the null hypothesis that the bias of z is equal to zero, see whether V is a valid estimator of the variance of z , or whether the confidence intervals based on V are valid statistical inferences. Finally, the variance S_{zz} , the corresponding mean square $(S_{zz} + (\bar{z} - \mu)^2)$ or their square roots can be used to compare the efficiency of various combinations of sampling methods and estimation procedures.

The analysis of the simulation results led to the several important conclusions listed in the previous section. In particular one should note that (i) the ordinary least squares method is almost as good as the ordinary weighted least squares but they both underestimate the error of the estimators; the only exception being the case of $p=5$ (less than two trees, on the average per cluster) where the error of the OWLS (but not that of the OLS) estimators is properly evaluated by V , (ii) the modified least and weighted least squares estimators are somewhat less precise than the corresponding ordinary ones, when the number of sample clusters within a given stratum is sufficiently large, say greater than ten; the modified estimation approaches should never be used, however, when the number of sample clusters is small, say less than ten, (iii) the usual procedure of calculating the biomass regression function by combining the data of the trees from all strata (procedure 1) is dangerous; it may become an important source of bias, (iv) the best overall procedure in terms of bias and accuracy is that of constructing biomass regressions separately by stratum (procedure 2) and then applying the usual stratified sampling formulae to derive estimators and their error.

The surprising conclusion about the equal efficiency of least and weighted least squares method should be properly interpreted. They are equally good in terms of the bias and the precision of the estimator of the average biomass per unit area. If one wishes to calculate a biomass regression function with valid estimates of its precision (confidence and prediction in-

tervals) one must use the weighted least squares method.

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

- Arabatzis, A. A. Evaluating the error of tree biomass regressions by simulation; sample trees selected by stratified cluster sampling. M.S. thesis, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Cunia, T. Dummy variables and some of their uses in regression analysis. In: Proceedings of the June 1973 Meeting of IUFRO Subject Group S4.02, Nancy-France, Vol. 1, T. Cunia, K. Kuusela, and A. J. Nash (Eds.), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1973.
- Cunia, T. Evaluating errors of tree biomass regressions by simulation. In: Proceedings of the workshop on "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.; Gillespie, A. J. Cluster sampling and construction of biomass tables: results of a simulation study. In: Proceedings, third annual southern forest biomass workshop, March 12-14, 1985; University of Florida, Gainesville, FL; 1985.
- 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; 1983.
- 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, NC; 1984.
- Gillespie, A. J. Estimation of biomass tables by cluster sampling: a simulation study. M.S. thesis, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1985.
- Gillespie, A. J.; Cunia, T. Error of biomass regressions: sample trees selected by cluster sampling. In: Proceedings of the Workshop on "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.
- 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.
- Michelakackis, J.; Cunia, T. Error of biomass regressions: sample trees selected by double sampling. In: Proceedings of the workshop on "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.

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 $\hat{y} = a_1 + a_2 d^2 h$ and $\hat{h} = y_1 + y_2 d + y_3 d^2$ 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 $[b]' = [b_1 b_2 b_3 b_4]$ of regression coefficients, when the vectors $[a]' = [a_1 a_2]$ and $[c]' = [c_1 c_2 c_3]$ 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

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

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 $y = a_1 + a_2 d^2 h$ 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

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

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, $\sqrt{V/S_{zz}}$, $(S_{zz})/\bar{z}$, $t = (\bar{z} - \mu)/\sqrt{S_{zz}/k}$ and the number of times μ fell within the 95 and 99 percent confidence intervals $(\bar{z} \pm t\sqrt{V})$ 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 $(\bar{z} - \mu)$ 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 $\sqrt{V}/S_{zz}$.

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 $\hat{y} = r_3(d)$ calculated from the data of the first phase sample only) is not as good as the double sampling estimator z_2 (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

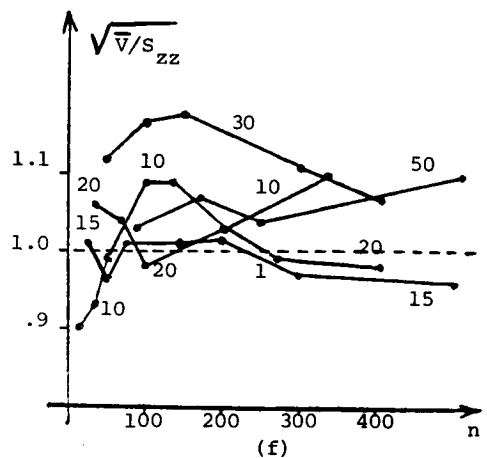
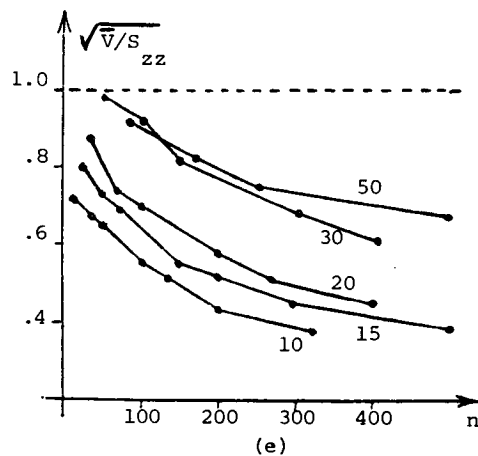
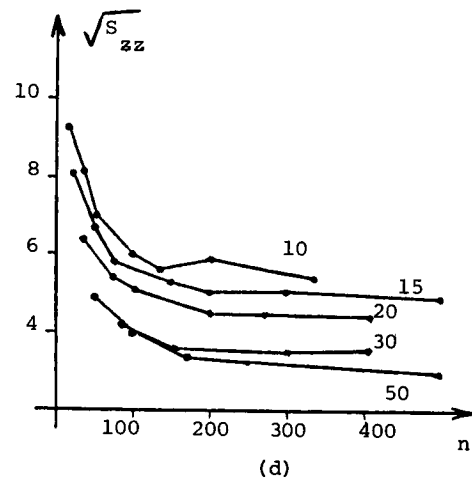
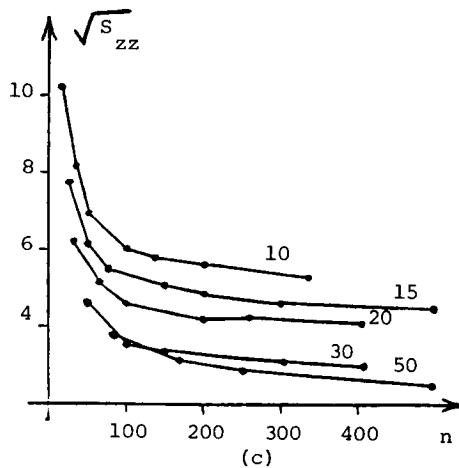
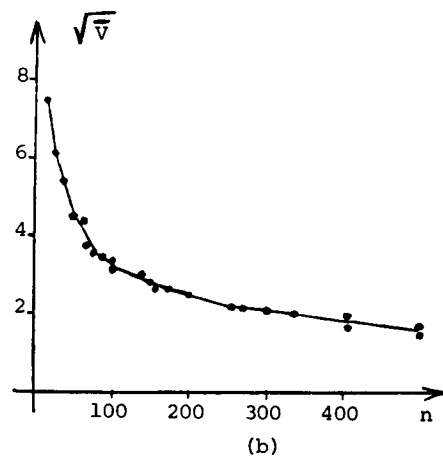
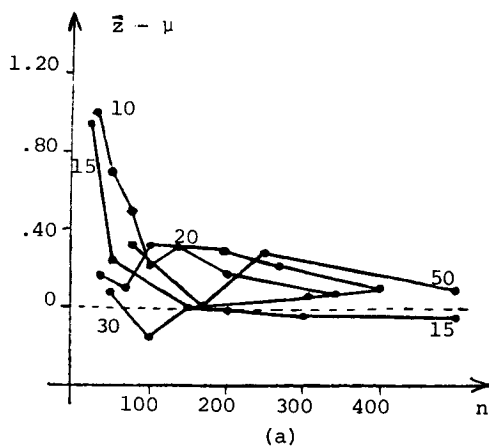


Figure 1 - Some of the relationships of $(\bar{z}-\mu)$, $\sqrt{s_{zz}}$, $\sqrt{v}$ and $\sqrt{v/s_{zz}}$ with average sample size n of the first phase sample, for the data of Table 1. The number 10, 15, 20, 30 and 50 alongside a curve shows the number m of sample clusters of the first phase sample.

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 $\hat{V}$ 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 Michelakakis 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 $\sqrt{V}/\sqrt{S_{zz}}$ 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 $\sqrt{V/S_{zz}}$ 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	$\sqrt{V/S_{zz}}$	$\bar{z}-\mu$	Estimator $\sqrt{S_{zz}}$	t	$\sqrt{V/S_{zz}}$	$\bar{z}-\mu$	Estimators $\sqrt{S_{zz}}$	t	$\sqrt{V/S_{zz}}$
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 Michelakakis 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, $\hat{y} = r_3(d)$ 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.

^{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.

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.

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 $\hat{v}(\bar{Y}_1 - \bar{Y}..) =$	$(1 - \frac{n}{N}) \frac{s_{Y_1}^2}{n}$		$+ \frac{1}{n} \sum_{i=1}^n T_i^2 \sigma_i^2 \left[\frac{T_i}{nt_i} + \frac{T_i (\bar{X}_i - \bar{X})^2}{n(t_i - 1)s_{x_i}^2} - \frac{1}{N} \right]$
2 $\hat{v}(\bar{Y}_2 - \bar{Y}..) =$	$(1 - \frac{n}{N}) \frac{s_{Y_2}^2}{n}$	$+ (1 - \frac{m}{n}) \frac{s_d^2}{m}$	$+ \frac{1}{m} \sum_{i=1}^m T_i^2 \sigma_i^2 \left[\frac{T_i}{mt_i} + \frac{T_i (\bar{X}_i - \bar{X})^2}{m(t_i - 1)s_{x_i}^2} - \frac{1}{N} \right]$
3 $\hat{v}(\bar{Y}_3 - \bar{Y}..) =$	$(1 - \frac{n}{N}) \frac{s_{Y_3}^2}{n}$		$+ \frac{1}{n} \sum_{i=1}^n T_i^2 \sigma_i^2 \left[\frac{T_i}{mt} + \frac{T_i (\bar{X} - \bar{X})^2}{(mt - 1)s_x^2} - \frac{1}{N} \right]$

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.

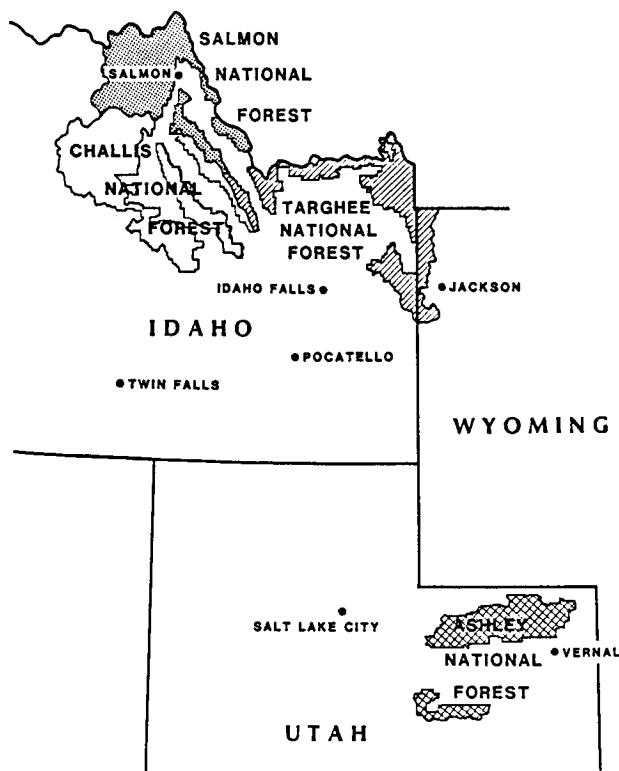


Figure 1.--Area map of the Salmon, Challis, Targhee, and Ashley National Forests.

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.

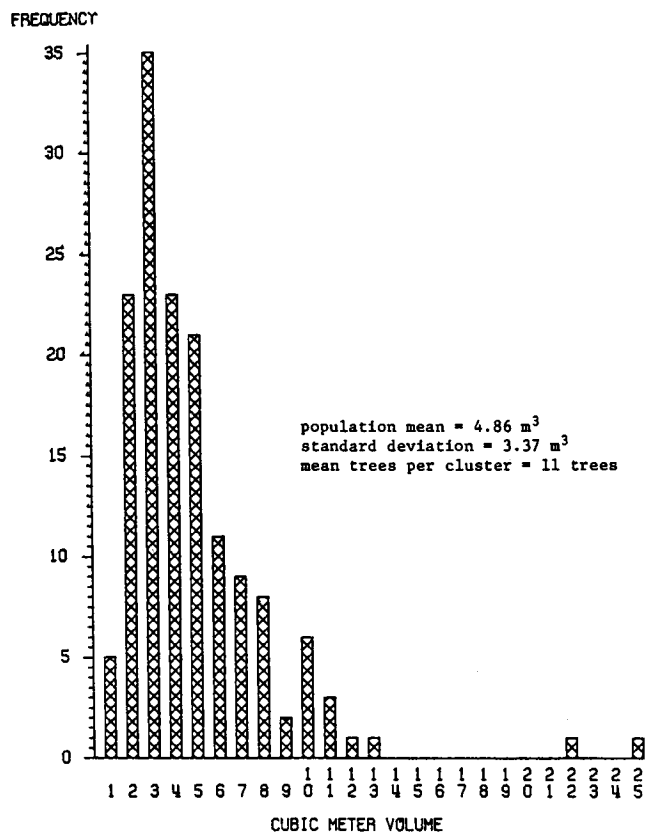
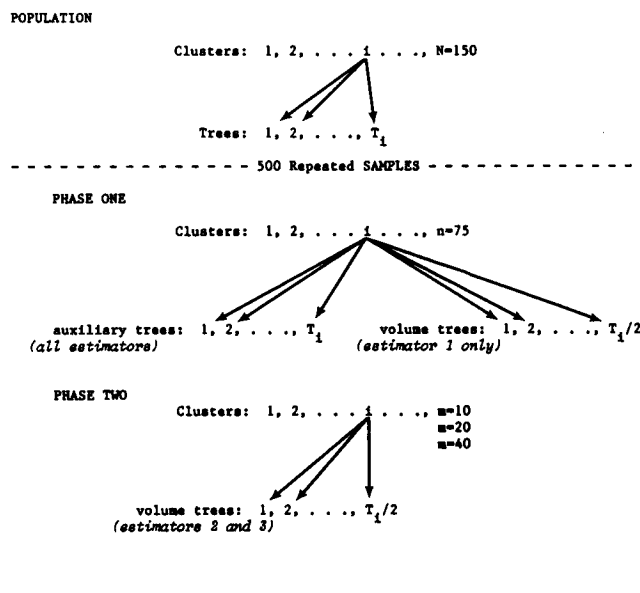


Figure 2.--Frequency distribution of cubic meter volume per cluster for the population.

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:

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

where

$\hat{\bar{y}}_{ij}$ = 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 ($m = 10, 20$, or 40). Bias for each estimator was determined by comparing the population mean (4.856 m^3 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).

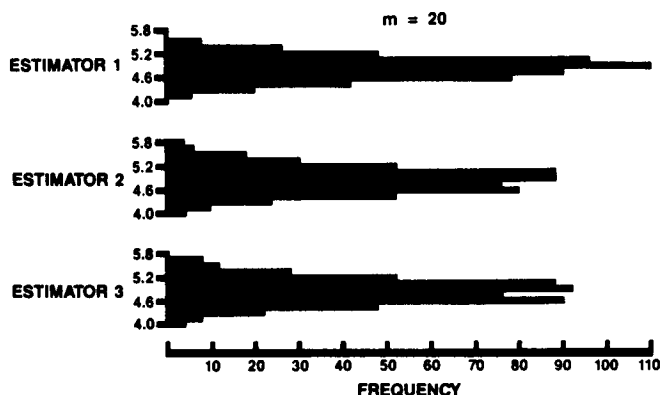


Figure 3.--Frequency distributions of sample means for m=20.

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	<u>Sample sizes</u>		Bias ^{a/}
	n	m	
			<u>Percent</u>
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.

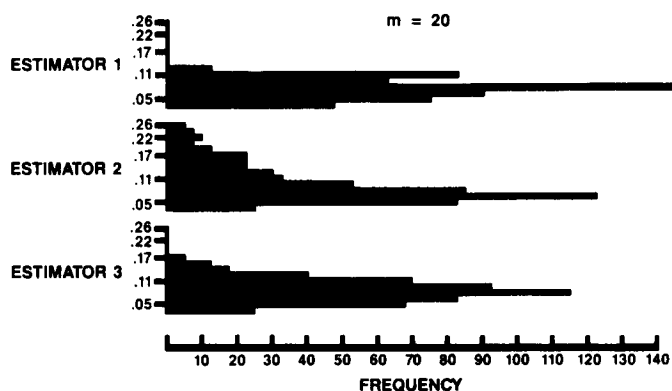


Figure 4.--Frequency distributions of estimator sample variances for m=20. (Note the nonuniform vertical scale.)

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

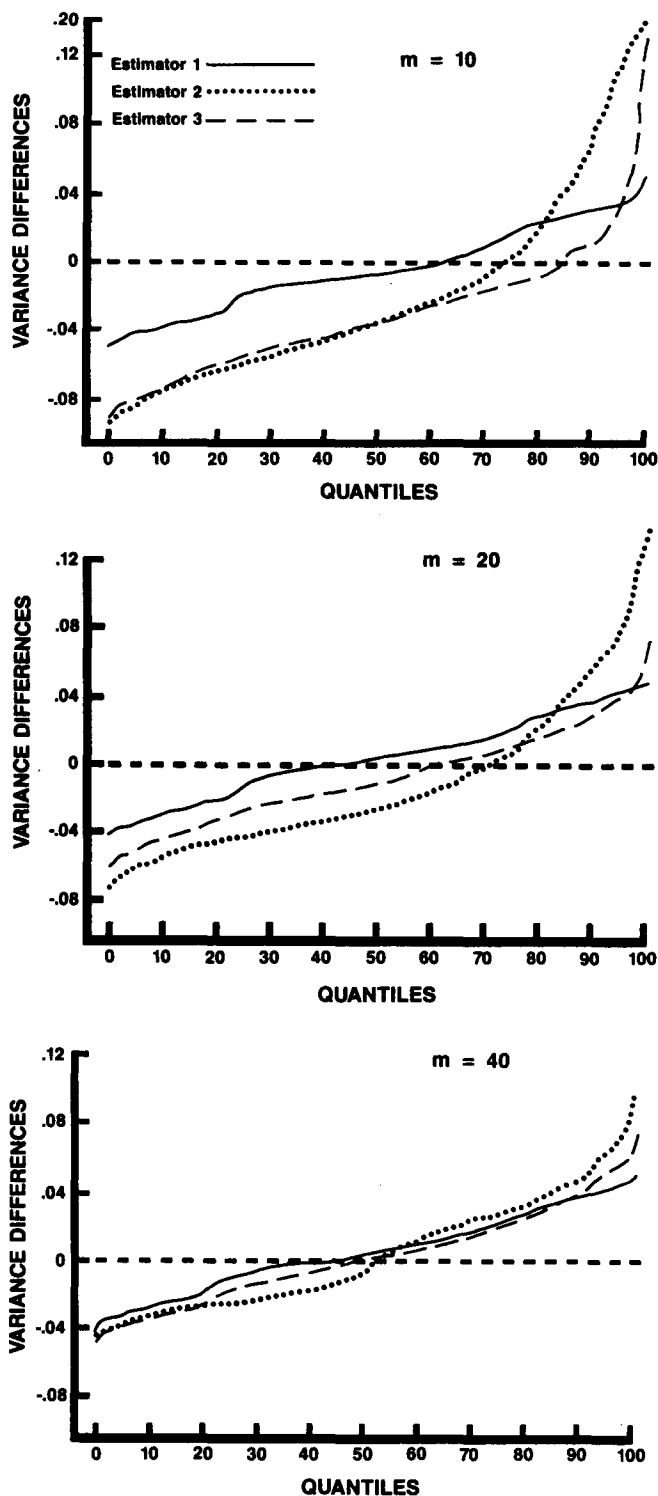


Figure 5.--Difference between the sample variance of the estimator and its corresponding simulation variance for quantiles of the sample variance distribution for m equal to 10, 20, and 40. (Note the nonuniform vertical scale for $m=10$.)

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:

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

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)
- $df_{1ij} = \begin{cases} n-1 & \text{for } i=1 \\ m-1 & \text{for } i=2,3 \end{cases}$
- $df_{2ij} = m-1 \text{ for all } i$
- $df_{3ij} = \begin{cases} n-1 & \text{for } i=1 \\ m-1 & \text{for } i=2,3. \end{cases}$

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 $\frac{n}{m}$		Degrees of freedom ^{a/}	Confidence intervals	
				Simulation variance	Sample variance
				Percent containing population mean	
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.

Kemphorne, 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.

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

$$\hat{\alpha}_1 = \bar{y}_1 - \hat{\beta}_1 \bar{x}_1$$

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

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{t_1} (x_{1j} - \bar{x}_1) y_{1j}}{\sum_{j=1}^{t_1} (x_{1j} - \bar{x}_1)^2}$$

$$d_1 = (\hat{\alpha}_1 T_1 + \hat{\beta}_1 X_1 - R X_1) (\bar{X} / \bar{X})$$

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

D^2H = 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

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

$$\hat{\sigma}^2 = \sum_{i=1}^m \sum_{j=1}^{t_i} \frac{(y_{1j} - \hat{\alpha}_1 - \hat{\beta}_1 x_{1j})^2}{(mt_i - 2)}$$

$$\hat{\sigma}_1^2 = \sum_{j=1}^{t_1} \frac{(y_{1j} - \hat{\alpha}_1 - \hat{\beta}_1 x_{1j})^2}{(t_1 - 2)}$$

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

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

$$s_{x_1}^2 = \sum_{j=1}^{t_1} \frac{(x_{1j} - \bar{x}_1)^2}{(t_1 - 1)}$$

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

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

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

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_{1j} = D^2H for j th tree of i th cluster

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

$$\bar{x}_1 = \sum_{j=1}^{t_1} \frac{x_{1j}}{t_1}$$

$$X_1 = \sum_{j=1}^{T_1} x_{1j}$$

$$\bar{X}_1 = \sum_{j=1}^{T_1} \frac{x_{1j}}{T_1}$$

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

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

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

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

y_{1j} = observation for j th element of i th cluster (volume of a tree)

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

$$\bar{y}_1 = \sum_{j=1}^{t_1} \frac{y_{1j}}{t_1}$$

$$\hat{Y}_{1i} = \hat{\alpha}_1 T_i + \hat{\beta}_1 X_i$$

$$\hat{Y}_{2i} = (\hat{\alpha}_1 T_i + \hat{\beta}_1 X_i) (\bar{X} / \bar{X})$$

$$\hat{Y}_{3i} = \hat{\alpha}_1 T_i + \hat{\beta}_1 X_i$$

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

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

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

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

$$\hat{\bar{Y}} = \frac{1}{m} \sum_{i=1}^m \hat{\alpha}_1 T_i + \hat{\beta}_1 X_i$$

"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 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).

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

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

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

$$V = (b_0 + b_2a_1) + (b_1 - b_2)X + \epsilon \text{ for } X < a_1, \\ = b_0 + b_1X + \epsilon \text{ for } a_1 \leq X \leq a_2 \\ = (b_0 + b_3a_2) + (b_1 - b_3)X + \epsilon \text{ for } a_2 \leq X \\ \epsilon \sim N(0, \sigma^2X^k) \quad (4)$$

where

$V = V_t$ or V_m

$X = D^2H_t$ or D^2H_4

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

b_1 = 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 D^2H intervals to calculate variance in volume as a function of tree size (D^2H). This variance is a function of interval length because the variance increases with the range of D^2H values in any one interval. Also, there is only one data point per interval (mean D^2H and $\text{Var}(V)$ for that interval) that can be used in the regression model (6) for estimating k . These factors encourage use of short D^2H 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 $\text{Var}(V)$ and tree size. We recommend sample sizes of 12-20 trees in each D^2H interval for data sets that are comparable to ours in size and number of D^2H 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 (a_1 and a_2) 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 D^2H units from 1,000 to 40,000, subject to $a_1 \leq (a_2 - 1,000)$. 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	.2345
	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 join-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 models⁰ (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 hypothesis 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 forms with both H_4 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_t 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_t 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 $V_t = f(D^2H_t)$, 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>						
$V_m = f(D^2H_4)$, $k=1.79$, $n=959$						
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
$V_t = f(D^2H_4)$, $k=1.43$, $n=959$						
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 Unweighted (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_4 --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
		1	2	3	4		sample size (n)
<u>LOBLOLLY PINE</u>							
<u>V_m = f(D²H₄) 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	-16.45 ^d	1.68 ^d	-2.46	10.04	40
	1,420	3.52 ^d	-7.22 ^d	2.26 ^d	0.29	9.83	59
Sawtimber	2,350	1.51	-5.95 ^d	0.08	-1.05	8.24	45
	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 ^d	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 ^d	1.77 ^d	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	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 ^d	2.63	13.29	42
	50,000	-8.55 ^d	-2.53	-3.54 ^d	-2.61	10.59	55
	<u>V_t = f(D²H₄), 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 ^d	-8.72	20.84	40
	1,420	10.35	-9.27 ^d	9.10 ^d	2.80	19.06	51
Sawtimber	2,350	-0.08	-10.97 ^d	-2.04	-4.41	18.67	45
	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 ^d	-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	
		1	2	3	4		(n)	
<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	22.55 ^d	7.82	105.11	45	
11,600		27.68 ^d	20.70 ^d	32.65 ^d	14.46 ^d	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 ^d	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	
	4,400	-1.80	-1.67	-0.71	-0.19	15.32	69	
	Sawtimber	5,300	-1.39	-1.59	-0.28 ^d	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 ^d	-0.49 ^d	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 ^d	0.10 ^d	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	
	5,350	0.59	0.66	-0.59	0.18	7.28	49	
Sawtimber	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 ² H class		Mean for each model form ^a				Var(R) ^b	Test sample size
		1	2	3	4		(n)
V _t =f(D ² H ₄), 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 ^d	-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 ² H _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 ^d	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 ^d	0.79 ^d	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 ² H _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
Sawtimber	5,350	-6.61 ^d	-4.06	-4.95	-5.02	43.55	50
	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.