



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

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

(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

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

(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 $\mu =$ 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

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

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,

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

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

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

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

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

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

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

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

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

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

where

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

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

By expressing \hat{v}_h as

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

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

by evaluating $\sum \hat{v}^2$ 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

$$y_{hk} = b_1 + b_2 d_{hk} + b_3 d_{hk}^2$$

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, ..., 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, ..., 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 \hat{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, ..., 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)

$$S_{ww}^{(3)} = \Sigma \Sigma b_i b_j S_{z_i z_j} = 11292.170$$

= 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

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

and

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

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

$$[S_{zz}^{ij}] = \begin{bmatrix} S_{z_1i z_1j} & S_{z_1i z_2j} & S_{z_1i z_3j} \\ S_{z_2i z_1j} & S_{z_2i z_2j} & S_{z_2i z_3j} \\ S_{z_3i z_1j} & S_{z_3i z_2j} & S_{z_3i z_3j} \end{bmatrix}$$

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

$$s_{h1} = \Sigma(1/a_{hk})$$

$$s_{h2} = \Sigma(d_{hk}/a_{hk}), \text{ and}$$

$$s_{h3} = \Sigma(d_{hk}^2/a_{hk})$$

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

$$a_{hk} = \pi (d_{hk}/12)^2 \text{ square feet}/4c(\text{square feet/acre})$$

$$= (\pi d_{hk}^2/576c) \text{ acres}$$

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

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

$$= (\pi d_{hk}^2/4c) \text{ hectares}$$

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

$$a_{hk} = \frac{\pi (d_{hk} \text{ centimeters}/(100 \text{ centimeters/meter}))^2}{4c(\text{square meters/hectare})}$$

$$= \frac{\pi d_{hk}^2 \text{ square meters}/10000}{4c \text{ square meters/hectare}}$$

$$= (\pi d_{hk}^2/40000c) \text{ hectares}$$

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

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

= volume on a per acre basis at the sample point h

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

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,

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

+ 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 $n_c = 353$ 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

$$s_{h1} = \Sigma (1/10a_{hk}) = (\text{number of trees}) \text{ per acre at cluster } h$$

$$s_{h2} = \Sigma (d_{hk}/10a_{hk}) = (\text{sum of tree diameters}) \text{ per acre at cluster } h$$

and

$$s_{h3} = \Sigma (d_{hk}^2/10a_{hk}) = (\text{sum of squared diameters}) \text{ per acre at cluster } h,$$

where Σ is taken over the trees k, k = 1, 2, ..., n_h counted at the cluster h of size n_h, divisor a_{hk} for a diameter d_{hk} is equal to (πd_{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

$$s_{h1} = \Sigma (576c/10\pi d_{hk}^2) = 687.54935 \Sigma (1/d_{hk}^2)$$

$$s_{h2} = \Sigma (576c/10\pi d_{hk}) = 687.54935 \Sigma (1/d_{hk}) \text{ and}$$

$$s_{h3} = \Sigma (576c/10\pi) = 687.54935 n_h$$

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

$$s_{11} = 687.54935(1/(16.3)^2 + 1/(5.0)^2 + \dots + 1/(8.3)^2) = 275.67758$$

$$s_{12} = 687.54935(1/16.3 + 1/5.0 + \dots + 1/8.3) = 2171.6585$$

$$s_{13} = (687.54935)(1 + 1 + \dots + 1) = 18563.833$$

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

$$\Sigma s_{h1} = 358160.27, \Sigma s_{h2} = 3077217.2, \Sigma s_{h3} = 30776084,$$

$$\Sigma s_{h1}^2 = 71609667, \Sigma s_{h1}s_{h2} = 601337690, \Sigma s_{h1}s_{h3} = 5822135900,$$

$$\Sigma s_{h2}^2 = 5164432500, \Sigma s_{h2}s_{h3} = 51276564000, \Sigma s_{h3}^2 = 524756860000$$

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

easy to calculate

$$\bar{s}_1 = (358160.27)/(4313) = 83.042029$$

= estimate of the average number of trees per acre

$$S_{11} = (71609667 - (358160.27)^2/4313)/(4312) = 9709.4878$$

= estimate of the variance of the cluster number of trees per acre

$$S_{12} = (601337690 - (358160.27)(3077217.2)/4313)/(4312) = 80194.650$$

= estimate of the covariance of the two cluster variables s_{h1} (the number of trees per acre) and s_{h2} (the sum of tree diameters per acre)

and similarly

$$\bar{s}_2 = 713.47489 = \text{estimate of the average sum of tree diameters per acre}$$

$$\bar{s}_3 = 7135.6560 = \text{estimate of the average sum of squared tree diameters per acre}$$

$$S_{22} = 688523.95 = \text{estimate of the variance of } s_{h2} \text{ (sum of diameters per acre)}$$

$$S_{13} = 757520.27 = \text{estimate of the covariance of } s_{h1} \text{ and } s_{h3}$$

$$S_{33} = 70767465 = \text{estimate of the variance of } s_{h3} \text{ (sum of squared diameters per acre)}$$

and

$$S_{23} = 6799304.4 = \text{estimate of the covariance of } s_{h2} \text{ and } s_{h3}$$

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

$$w = [b]'[z] = 74807.927 \text{ pounds per acre } \mu$$

= estimate of the average biomass per acre μ

$$S_{ww}^{(1)} = [b]'[S_{zz}][b] = 1841261.6$$

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

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

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

$$S_{ww}^{(3)} = \sum \sum S_{b_{ij}} S_{z_{ij}} = 1119.6811$$

= 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

$$\begin{aligned} \partial L / \partial n_{ij} &= \partial C_1 / \partial n_{ij} + \partial C_2 / \partial n_{ij} + \lambda \partial V_1 / \partial n_{ij} \\ &\quad + \lambda \partial V_2 / \partial n_{ij} = 0 \end{aligned}$$

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, $s_i = (x_i/a)$ where "a" is the plot area and Σ 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 $x_i = d$ (tree diameter) then $s_i = (\text{sum of the tree diameters in the plot})/(.20)$. 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

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

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

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

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_t^{**} 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_1 (defined as x_1 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

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

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

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

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

and

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

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

$$n_t = (6848.7313)(1517018.1)/(20,000,000) = 519.48 = 519$$

(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

$$n_t = (120000)(6848.7313)/(1517018.1) = 541.75 = 542$$

(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 $[S_{bz}^{ij}]$ 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

$$(2) S_{ww} = [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z] = [B]'[S_{ZZ}][B] + [z^1]'[S_{bb}^{11}][z^1] + [z^2]'[S_{bb}^{22}][z^2] + \dots + [z^m]'[S_{bb}^{mm}][z^m]$$

where \hat{w} = estimator of the variance of w

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

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

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

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

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

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

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

$$[\beta]' = [[\beta^1]', [\beta^2]', \dots, [\beta^q]']$$

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

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 $[\beta]$, $[\mu_z^i]$, $[\sigma]$, $[\theta^{ii}]$, 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 $[\beta]$, $[z^i]$ 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,

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

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

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 $[\theta^{ii}] = S_{uu|v} [T]^{-1} = n_i [S^{ii}]_{bb}$ 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 $Q_p, Q_1, Q_2, \dots, Q_q$ 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 $C^* = \$128230$

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

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

$$n_2 = (128230)(2236.9665)/(1480828.7) = 194$$

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

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 $n_p = 791, n_1 = 129, n_2 = 183, n_3 = 218$ 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.