

TABLE OF CONTENTS

PART I: TUTORIAL PAPERS

Combining the Error of Sample Plots and Biomass Regressions

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

Error of Biomass Regressions

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

Error of Sample Plots

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

PART II: RESEARCH PAPERS

Biomass Regressions and Measurement Error

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

Biomass of Forest Understory Vegetation

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

TABLE OF CONTENTS

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

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

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

Biomass Studies Outside the United States

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

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

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

¹Contributed paper, not presented at the workshop.

TUTORIAL PAPERS

Error of Sample Plots

ON THE ERROR OF FOREST INVENTORY ESTIMATES:
 STRATIFIED SAMPLING AND DOUBLE SAMPLING FOR
 STRATIFICATION^{1/}

Tiberius Cunia

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

When the design of a forest inventory system is stratified sampling and the stratified sampling formulae are used to calculate the error of the estimate of the average biomass per acre, it is common to ignore the error of the biomass regression function used to estimate the sample tree and plot biomass. The approach described in an earlier paper by Cunia to combine the error of biomass regressions with the error from sample plots, when estimates of average biomass per acre are calculated, is extended from simple random to stratified random sampling. As sometimes the stratum sizes are not known but estimated by a second sample we have also considered the double sampling for stratification design.

Introduction

Cunia (1986a) has proposed an approach to combine the error of the biomass regression function with the error of the forest inventory sample plots (or Bitterlich sample points) when inferences are made about the reliability of the biomass estimates per unit area. This approach requires that the estimators be of the form

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

where [b] is the vector of the coefficients of the biomass regression and [z] is a vector of statistics calculated from the data of the sample plots and points. We implicitly assume here that (i) the true regression function of tree biomass on $[x]' = [x_1 \ x_2 \ \dots \ x_m]$ is of the linear form $\hat{y} = [\beta]'[x]$, (ii) the vector [z] is defined so that, the product $[\beta]'[u_z]$, where $[u_z]$ is the expected value of [z], is close, if not identically equal to the parameter of interest μ and (iii) the vectors [b] and [z] are statistically independent. The variance of w is estimated by the approximate formula

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

where $[S_{zz}]$ and $[S_{bb}]$ are the estimates of the covariance matrices of [z] and [b] respectively.

^{1/}Paper based on a set of lecture notes "On the error of biomass estimates in forest inventories: Part 2: The error component from sample plots". Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

In this formula, the first component of S_{ww} may be viewed as an expression of the error due to the sample plots and the second component may be viewed as an expression of the error due to the biomass regression function.

The definition of [z] depends on (i) the sampling design by which the plots or points are selected, (ii) the specific parameter μ one wishes to estimate and (iii) the definition of the independent variables x_1, x_2, \dots, x_m of the regression function. In his paper, Cunia (1986a) makes the assumptions that (i) the sample plots are selected by simple random sampling (or by a systematic sampling method that is equivalent to simple random sampling), (ii) the parameter to estimate is μ = average biomass per acre, and (iii) the definition of the statistics z is based on the plot variables s_1, s_2, \dots, s_m defined as the averages of x_1, x_2, \dots, x_m expressed on a "per acre" basis.

For example, let us assume that the estimate of the regression function of the tree biomass y on the tree diameter d is of the parabolic form

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

where the definition of x_1, x_2 and x_3 is obvious. Then, for Σ meaning summation over all the trees of a plot, the plot variables are defined as

- $s_1 = (\Sigma x_1)/\text{plot area}$
= number of trees per acre
- $s_2 = (\Sigma x_2)/\text{plot area}$
= sum of diameters per acre, and
- $s_3 = (\Sigma x_3)/\text{plot area}$
= sum of squared diameters per acre.

When, instead of a sample plot we have a Bitterlich sample point, there are similar definitions for the point variables s_1, s_2 and s_3 ; these are not repeated here. For more details, the interested reader is referred to Cunia (1986a).

Let us assume now that the plots (or points) are selected by simple random sampling and the parameter to estimate is μ , the average biomass per acre. Then, the elements of [z] are defined as

- $z_1 = \bar{s}_1 = \Sigma s_1/n$ = estimate of the average "number of trees" per acre,
- $z_2 = \bar{s}_2 = \Sigma s_2/n$ = estimate of the average "sum of diameters" per acre, and
- $z_3 = \bar{s}_3 = \Sigma s_3/n$ = estimate of the average "sum of squared diameters" per acre

where Σ means summation over the n plots (or points) in the sample.

If the sample covariance of s_i and $s_j, i, j = 1, 2, 3$ is defined as usual by the formula

$$S_{s_i s_j} = \Sigma (s_i - \bar{s}_i)(s_j - \bar{s}_j)/(n-1)$$

and the estimate of the covariance matrix of s_1 , s_2 and s_3 is denoted by $[S_{SS}]$, then, the estimate of the covariance matrix of $[z]$ is

$$[S_{ZZ}] = [S_{SS}]/n$$

Consequently, the estimate of μ and that of its error is

$$w = [b]'[z] \text{ and } S_{ww} = [b]'[S_{ZZ}][b] + [z]'[S_{bb}][z]$$

where $[S_{bb}]$ is the estimate of the covariance matrix of $[b]$.

It is the objective of this paper to extend the approach described by Cunia (1986a) from simple random to stratified random sampling; that is, to define a stratified sampling estimator of the average biomass per acre μ and to propose a formula to estimate its error. As sometimes the stratum size is not known without error we shall also consider the case where the size of strata are estimated from a second sample. We shall assume that the reader is familiar with the previous paper by Cunia (1986a) and he understands the definition of the plot variables s_1, s_2, \dots, s_m . We shall also assume that the reader is familiar with the theory of the methods of stratified random sampling and double sampling for stratification as described, among others, by Cochran (1977).

Stratified Sampling Applied to Forest Inventory

Consider a forest area subdivided into L non-overlapping and exhaustive strata, where $A_h =$ area (in acres) of stratum h , $A =$ area of the entire forest and $Q_h = A_h/A =$ relative size of stratum h . A statistically independent simple random sample of plots of size $n_h > 2$ is selected from each stratum h and, for notational convenience, we shall assume that n_h is small with respect to the size of the stratum h so that, the effect of the finite population correction factor can be ignored. If, for Σ meaning summation over the n_h plots of stratum h , we define

$v_{hk} =$ biomass "per acre" of plot k in stratum h

$\bar{v}_h = (\Sigma v_{hk})/n_h =$ estimator of the average biomass per acre of stratum h , and

$S_{vv}^{hh} = \Sigma (v_{hk} - \bar{v}_h)^2 / (n_h - 1) =$ estimator of the variance of v_{hk} within stratum h

then, for Σ meaning summation over the L strata,

$\bar{v} = \Sigma Q_h \bar{v}_h =$ stratified sample mean
 = estimator of the mean biomass per acre for the entire forest area, and

$S_{\bar{v}\bar{v}} = \Sigma Q_h^2 S_{vv}^{hh} / n_h =$ estimator of the variance of \bar{v}

When the variance of \bar{v} above is calculated, the error of the biomass regression function (used to calculate v_{hk}) is ignored. To take it

into account we shall write w as the estimator

$$\begin{aligned} w &= \Sigma Q_h \bar{v}_h = \Sigma Q_h (\Sigma v_{hk} / n_h) \\ &= \Sigma Q_h (b_1 \Sigma s_{1hk} + b_2 \Sigma s_{2hk} + \dots + b_m \Sigma s_{mhk}) / n_h \\ &= \Sigma Q_h (b_1 \bar{s}_{1h} + b_2 \bar{s}_{2h} + \dots + b_m \bar{s}_{mh}) \\ &= b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]'[z] \end{aligned}$$

where the first Σ means summation over stratum h , the second Σ means summation over the values s_{hk} of plots k within stratum h , and $z_i = \Sigma Q_h \bar{s}_{ih} =$ stratified mean of the variables $s_i, i=1,2,\dots, m$.

The estimator of the variance of z_i is given by the usual stratified sampling formula

$$S_{z_i z_i} = \Sigma Q_h^2 S_{ss}^{hh} / n_h$$

while the estimator of the covariance of z_i and z_j is a simple extension of the above formula, that is,

$$S_{z_i z_j} = \Sigma Q_h^2 S_{ss}^{hh} / n_h$$

It is more convenient to express these results in a matrix notation as

$$[\bar{s}^h]' = [\bar{s}_{1h} \quad \bar{s}_{2h} \quad \dots \quad \bar{s}_{mh}]$$

$$[S_{SS}^{hh}] = \begin{bmatrix} S_{ss}^{hh} & S_{ss}^{hh} & \dots & S_{ss}^{hh} \\ S_{ss}^{hh} & S_{ss}^{hh} & \dots & S_{ss}^{hh} \\ \vdots & \vdots & \ddots & \vdots \\ S_{ss}^{hh} & S_{ss}^{hh} & \dots & S_{ss}^{hh} \end{bmatrix}$$

and

$$[S_{ZZ}] = \Sigma Q_h^2 [S_{SS}^{hh}] / n_h$$

If the estimates of the average biomass per acre for stratum h is required, then

$w_h = [b]'[z^h] =$ estimator of the average biomass per acre μ_h of stratum h

and $S_{w_h w_h} = [b]'[S_{ZZ}][b] + [z^h]'[S_{bb}][z^h]$
 = estimator of the variance of w_h

where

$$[z^h] = [\bar{s}^h] \text{ and } [S_{ZZ}^{hh}] = [S_{SS}^{hh}] / n_h$$

If the estimate of the average biomass per acre for the entire forest area is required then

$w = [b]'[z] =$ estimator of the average biomass per acre

and $S_{ww} = [b]'[S_{ZZ}][b] + [z]'[S_{bb}][z]$
 = estimator of the variance of w
 when $[z]$ and $[S_{ZZ}]$ have been defined above.

Example 1 - A forest area of 42336 acres is subdivided into three strata and we shall assume that the size of stratum 1, 2 and 3 is known to be equal, without error, to 17164, 19056 and 6116 acres respectively. Three statistically indepen-

dent samples, one for each stratum, were selected by simple random sampling. There are 82, 112 and 41 sample plots respectively in the samples of stratum 1, 2 and 3. To calculate the biomass we have used the regression function

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

with the obvious definitions for x_1 , x_2 and x_3 , where y = above ground biomass (pounds of green weight) and d = tree diameter (inches). The original data of the 353 sample trees and the calculations to determine, by weighted least squares, the vector $[b]$ and covariance matrix $[S_{bb}]$ are given in Cunia (1986b). For the convenience of the reader, $[b]$ and $[S_{bb}]$ are shown in Table 1.

The use of this biomass regression function and the fact that the trees are selected in clusters (plots) require the calculation of the three plot variables defined as

$$s_1 = \sum x_1 / a = \text{number of trees per acre}$$

$$s_2 = \sum x_2 / a = \text{sum of diameters per acre}$$

$$s_3 = \sum x_3 / a = \text{sum of squared diameters per acre}$$

The individual values of the (82 + 112 + 41) = 235 one tenth acre sample plots are given in Cunia (1986c) and the summary statistics by the stratum needed here, that is, the vectors $[\bar{s}^h]$ of mean values and covariance matrices $[S_{ss}^{hh}]$ for $h = 1, 2, 3$ are given in Table 1.

Using the relationships $[z^h] = [\bar{s}^h]$ and $[S_{zz}^{hh}] = [S_{ss}^{hh}] / n_h$ we can calculate the estimates w_h of the mean biomass per acre and their variances for each stratum h . For convenience, the matrices $[S_{zz}^{hh}]$ are shown in Table 1. The reader can verify that

$$w_1 = [b]'[z^1] = 37427.300 \text{ pounds} = \text{estimate of the mean biomass per acre } \mu_1 \text{ of stratum 1}$$

$$w_2 = [b]'[z^2] = 112268.69 \text{ pounds} = \text{estimate of the mean biomass per acre } \mu_2 \text{ of stratum 2}$$

$$w_3 = [b]'[z^3] = 225726.36 \text{ pounds} = \text{estimate of the mean biomass per acre } \mu_3 \text{ of stratum 3}$$

$$\begin{aligned} S_{w_1 w_1} &= [b]'[S_{zz}^{11}][b] + [z^1]'[S_{bb}][z^1] \\ &= 13014675 + 3221354 = 16236029 \\ &= \text{estimate of the variance of } w_1 \end{aligned}$$

$$\begin{aligned} S_{w_2 w_2} &= [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2] \\ &= 13124928 + 12147386 = 25272314 \\ &= \text{estimate of the variance of } w_2 \end{aligned}$$

$$\begin{aligned} S_{w_3 w_3} &= [b]'[S_{zz}^{33}][b] + [z^3]'[S_{bb}][z^3] \\ &= 53743500 + 49591085 = 103334585 \\ &= \text{estimate of the variance of } w_3 \end{aligned}$$

Table 1 - The basic statistics $[b]$, $[S_{bb}]$, $[\bar{s}^h] = [z^h]$, $[z]$, $[S_{ss}^{hh}]$, $[S_{zz}^{hh}]$ and $[S_{zz}]$ of Example 1.

$[b]'$	$\begin{bmatrix} 5.1818118 & -25.653078 & 12.988357 \end{bmatrix}$
$[S_{bb}]$	$\begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$
$[\bar{s}^1]'$	$\begin{bmatrix} 142.19512 & 743.47195 & 4293.2925 \end{bmatrix}$
$[\bar{s}^2]'$	$\begin{bmatrix} 277.58929 & 1647.1857 & 11786.376 \end{bmatrix}$
$[\bar{s}^3]'$	$\begin{bmatrix} 250.48780 & 1974.9829 & 21179.952 \end{bmatrix}$
$[z]'$	$\begin{bmatrix} 218.78217 & 1328.1538 & 10105.533 \end{bmatrix}$
$[S_{ss}^{11}]$	$\begin{bmatrix} 13753.147 & 69358.124 & 384753.02 \\ 69358.124 & 361014.14 & 2095548.5 \\ 384753.02 & 2095548.5 & 12995719 \end{bmatrix}$
$[S_{ss}^{22}]$	$\begin{bmatrix} 17946.388 & 77549.569 & 234084.91 \\ 77549.569 & 374176.41 & 1614218.3 \\ 234084.91 & 1614218.3 & 13563142 \end{bmatrix}$
$[S_{ss}^{33}]$	$\begin{bmatrix} 15249.756 & 69057.884 & 92042.322 \\ 69057.884 & 337241.41 & 973362.91 \\ 92042.322 & 973362.91 & 15624095 \end{bmatrix}$
$[S_{zz}^{11}]$	$\begin{bmatrix} 167.72130 & 845.83078 & 4692.1100 \\ 845.83078 & 4402.6115 & 25555.470 \\ 4692.1100 & 25555.470 & 158484.38 \end{bmatrix}$
$[S_{zz}^{22}]$	$\begin{bmatrix} 160.23561 & 692.40686 & 2090.0438 \\ 692.40686 & 3340.8608 & 14412.663 \\ 2090.0438 & 14412.663 & 121099.48 \end{bmatrix}$
$[S_{zz}^{33}]$	$\begin{bmatrix} 371.94527 & 1684.3386 & 2244.9347 \\ 1684.3386 & 8225.4001 & 23740.559 \\ 2244.9347 & 23740.559 & 381075.48 \end{bmatrix}$
$[S_{zz}]$	$\begin{bmatrix} 67.794455 & 314.46217 & 1241.5311 \\ 314.46217 & 1572.1751 & 7615.9949 \\ 1241.5311 & 7615.9949 & 58537.684 \end{bmatrix}$

To calculate the estimate of the mean biomass per acre μ for the entire forest area, we must calculate first the relative size of each stratum $h = 1, 2, 3$, that is, to calculate

$$Q_1 = A_1 / A = 17164 / 42336 = .40542328$$

$$Q_2 = A_2 / A = 19056 / 42336 = .45011338$$

$$Q_3 = A_3 / A = 6116 / 42336 = .14446334$$

Then, the numerical values of

$$[z] = Q_1 [z^1] + Q_2 [z^2] + Q_3 [z^3]$$

and

$$[S_{zz}] = (Q_1)^2 [S_{zz}^{11}] + (Q_2)^2 [S_{zz}^{22}] + (Q_3)^2 [S_{zz}^{33}]$$

are calculated and shown in Table 1. Note that $(Q_h)^2$ denotes here the square of Q_h .

Finally,

$w = [b]'[z] = 98316.722$
 = stratified sample estimate of the mean biomass per acre μ of the entire forest area, and

$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$
 = $5919942.4 + 9837039.2 = 15756982$
 = estimate of the variance of w .

The common procedure to calculate the stratified mean and its variance is to (i) apply the tree biomass regression function and calculate the biomass of each individual tree (ii) add the biomass of trees from a given sample plot to obtain the plot biomass, (iii) divide the plot biomass by plot area to obtain the plot biomass per acre v_{hk} , (iv) calculate the stratum average and variance by the formula $\bar{v}_h = \Sigma v_{hk}/n_h$ and $S_{vv}^{hh} = \Sigma (v_{hk} - \bar{v}_h)^2 / (n_h - 1)$ and finally (v) calculate the stratified mean and its variance by the formulae $\bar{v} = \Sigma Q_h \bar{v}_h$ and $S_{\bar{v}\bar{v}} = \Sigma Q_h^2 S_{vv}^{hh} / n_h$. Applied to our sample data the procedure yields the following statistics

$\Sigma v_{hk} = 3069038.6$	for stratum 1
$= 12574093.3$	for stratum 2
$= 9254780.8$	for stratum 3
$\Sigma (v_{hk})^2 = 2013092800$	for stratum 1
$= 1574846091000$	for stratum 2
$= 2177187330000$	for stratum 3

The individual plot values v_{hk} are listed in Cunia (1986c). Using these sums and sums of squares, one can calculate

$\bar{v}_h = \Sigma v_{hk} / n_h = 37427.300$	for stratum 1
$= 112268.69$	for stratum 2
$= 255726.36$	for stratum 3

and

$S_{\bar{v}\bar{v}}^{hh} = S_{vv}^{hh} / n_h = 13014675$	for stratum 1
$= 13124928$	for stratum 2
$= 53743500$	for stratum 3

As the reader can verify these are the same as the estimates w_1 , w_2 and w_3 found before and the corresponding first variance components, namely $[b]'[S_{zz}^{hh}][b]$ for $h = 1, 2, 3$.

Similarly

$\bar{v} = \Sigma Q_h \bar{v}_h = 98316.722 = w$

and

$S_{\bar{v}\bar{v}} = \Sigma Q_h^2 S_{\bar{v}\bar{v}}^{hh} = 5919942.4$
 = first component of the variance S_{ww} of w .

As expected, the variance of the mean biomass per acre is underestimated, when the common procedure is being applied; the second component associated with the biomass regression is being ignored. However, we did not expect to see this underestimation much more pronounced for the overall stratified mean. The percent of the total variance due to the error of biomass regression function is

$(100)(3221354)/(16236029) = 19.84$ for stratum 1
 $(100)(12147386)/(25272314) = 48.07$ for stratum 2
 $(100)(49591085)/(103334585) = 47.99$ for stratum 3

and

$(100)(9837039)/(15756982) = 62.43$ for the overall forest area.

This can be explained by the fact that stratification reduces the error due to sample plots but has no real effect on the error due to biomass regression functions.

The Case of Strata with Different Regressions

The approach of the previous section required that the same regression function be used in all strata. It is possible, however, for each stratum h to have its own regression. The regressions may or may not be statistically independent. We shall now present the approach to use when, in general, the regression functions of various strata are different and not statistically independent.

Let us denote by $[b^h]$ and $[S_{bb}^{hh}]$ the vector of regression coefficients and the corresponding covariance matrix of the biomass regression functions of stratum h . Because, for $h \neq k$, $[b^h]$ and $[b^k]$ may not be statistically independent, let us denote their covariance matrix by $[S_{bb}^{hk}]$. In this covariance matrix, the terms associated with the regression coefficient b_i^h are found on row i , while the terms associated with b_j^k are all found in column j . Of course, when $[b^h]$ and $[b^k]$ are independent $[S_{bb}^{hk}] = [0]$, and when $[b^h] = [b^k]$, we have $[S_{bb}^{hh}] = [S_{bb}^{kk}] = [S_{bb}^{hk}]$. Also, we must have $[S_{bb}^{kh}] = [S_{bb}^{hk}]'$.

We shall arrange all vectors $[b^h]$ in the following giant size vector $[B]$ and all covariance matrices $[S_{bb}^{hk}]$ in the following giant size covariance matrix $[S_{BB}]$, both of order mL , or order $(m_1 + m_2 + \dots + m_L)$ when the various regressions have different variables x ,

$$[B] = \begin{bmatrix} [b^1] \\ [b^2] \\ \cdot \\ \cdot \\ [b^L] \end{bmatrix} \quad \text{and} \quad [S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [S_{bb}^{12}] & \dots & [S_{bb}^{1L}] \\ [S_{bb}^{21}] & [S_{bb}^{22}] & \dots & [S_{bb}^{2L}] \\ \cdot & \cdot & \cdot & \cdot \\ [S_{bb}^{L1}] & [S_{bb}^{L2}] & \dots & [S_{bb}^{LL}] \end{bmatrix}$$

It is outside the scope of this paper to consider the problem of how to calculate the

covariance matrix $[S_{bb}^{hk}]$. Its value depends on (i) how the sample trees were selected from stratum h and k and (ii) how the data were analyzed and the regressions calculated. For the special case where the sample of trees from the two strata h and k are statistically independent, and some of the coefficients of $[b^h]$ and $[b^k]$ are thought to be common, one can use dummy variables techniques of the type suggested by Cunia (1973, 1986d). Otherwise, one may have to devise some generalized least squares techniques not generally available in textbooks or papers.

The estimates of the mean biomass per acre in stratum h and its variance can be calculated by the formulae of the previous section, namely

$$w_h = [b^h]'[z^h], \text{ and}$$

$$S_{w_h w_h} = [b^h]'[S_{zz}^{hh}][b^h] + [z^h]'[S_{bb}^{hh}][z^h]$$

The estimate of the overall forest area mean biomass per acre μ is calculated by the formula

$$w = Q_1 w_1 + Q_2 w_2 + \dots + Q_L w_L$$

$$= Q_1 [b^1]'[z^1] + Q_2 [b^2]'[z^2] + \dots + Q_L [b^L]'[z^L]$$

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

where the giant size vector $[Z]$ is defined as

$$[Z]' = \begin{bmatrix} Q_1 [z^1]' & Q_2 [z^2]' & \dots & Q_L [z^L]' \end{bmatrix}$$

Note that the vectors $[b]$ and $[z]$ above have superscripts, not exponents.

It remains now to show how to calculate the covariance matrix $[S_{ZZ}]$ of $[Z]$, defined here as

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

We have shown how to calculate the covariance matrix $[S_{zz}^{hh}]$ of $[z^h]$. Then, the covariance matrix of $Q_h [z^h]$ is simply the submatrix $[S_{ZZ}^{hh}] = (Q_h)^2 [S_{zz}^{hh}]$. To evaluate the submatrix $[S_{ZZ}^{hk}]$ we shall have first a close look at the covariance of $Q_h z_i^h$ and $Q_k z_j^k$. As z_i^h and z_j^k are statistically independent, their covariance is equal to zero and, thus, the covariance of $Q_h z_i^h$ and $Q_k z_j^k$ is also equal to zero. Consequently, for $h \neq k$, we have $[S_{ZZ}^{hk}] = [0]$, the zero matrix of order m.

The formulae of w and S_{ww} follow now as

$$w = [B]'[Z] = \text{estimator of the mean biomass per acre } \mu \text{ for the entire forest area,}$$

and

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

= estimator of the variance of w .

Special case 1 - The same regression function is used in all strata. Then

$$[B]' = \begin{bmatrix} [b]' & [b]' & \dots & [b]' \end{bmatrix}$$

$$[Z] = \begin{bmatrix} Q_1 [z^1]' & Q_2 [z^2]' & \dots & Q_L [z^L]' \end{bmatrix}$$

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

$$[S_{ZZ}] = \begin{bmatrix} [S_{ZZ}^{11}] & [0] & \dots & [0] \\ [0] & [S_{ZZ}^{22}] & \dots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \dots & [S_{ZZ}^{LL}] \end{bmatrix}$$

and it can be shown by complex and cumbersome algebraic manipulations that the formulae

$$w = [B]'[Z] \text{ and}$$

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

are identical to those obtained in the previous section. By using the numerical data of the illustrative example of the previous section, the reader can construct the giant size vectors and matrices defined here and verify that the two sets of formulae yield the same results.

Special case 2 - The regression functions of the various strata are all statistically independent, that is, for $h \neq k$, we have

$$[S_{bb}^{hk}] = [0]. \text{ Then}$$

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

$$= Q_1 [b^1]'[z^1] + Q_2 [b^2]'[z^2] + \dots + Q_L [b^L]'[z^L]$$

$$= Q_1 w_1 + Q_2 w_2 + \dots + Q_L w_L = \Sigma Q_h w_h$$

and

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

$$= \Sigma (Q_h)^2 \left([b^h]'[S_{zz}^{hh}][b^h] + [z^h]'[S_{bb}^{hh}][z^h] \right)$$

$$= \Sigma (Q_h)^2 S_{w_h w_h}$$

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

$w_h = [b^h]'[z^h]$. This formula makes sense, since, for $h \neq k$, the estimators w_h and w_k are statistically independent and by definition of w as $\Sigma Q_h w_h$, the variance formula follows immediately.

Double Sampling for Stratification

The approach outlined in the previous sec-

tions requires that the relative size Q_h be known without error for each stratum $h = 1, 2, \dots, L$. When it is not known, it is sometimes advantageous to select another sample for the sole purpose of estimating Q_h . The procedure is known as two-phase or double sampling for stratification.

More specifically, a first large sample of units is selected from the population of interest by simple random sampling and the units are classified by the stratum they happened to fall into. Let n' be the size of this sample and n'_h be the number of sample units that happened to fall in stratum h . We shall assume that n' is both, sufficiently small with respect to the population size (so that the effect of the finite population correction factor can be ignored) and sufficiently large (so that the probability of obtaining $n'_h = 0$, for some h , is approximately equal to zero). Then, it is known that

$Q_h = n'_h/n' =$ estimator of the relative size of stratum h

$S_{QQ}^{hh} = Q_h(1-Q_h)/n' =$ estimator of the variance of Q_h

and, for $h \neq k$,

$S_{QQ}^{hk} = -Q_h Q_k/n' =$ estimator of the covariance of Q_h and Q_k

Note that, for notational convenience we have used Q_h to denote both, statistic and parameter.

A second, stratified sample is then selected with n_h elements selected from each stratum h . The n_h elements may be a simple random subsample of the n'_h elements of the first sample above, or may be selected completely independent of them. The elements of this second sample are measured for the variables of interest s_1, s_2, \dots, s_m defined in the previous sections.

If we substitute the estimators Q_h for the corresponding true values in the formulae of the stratified sampling of the previous section, we obtain estimators of the mean volume per unit area, say μ and its variance. While the estimator of μ is valid, the estimator of its variance has a major drawback; it assumes that Q_h is known without error and, thus, the error of Q_h is simply ignored. We shall show now how to take this error component into account for both cases (i) when the second phase sample is a subsample of the first and (ii) when the first and second samples are statistically independent.

Case 1: Second Sample Is A Subsample of The First

The double sampling estimator can be written as

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

where

$$z_i = \sum Q_h \bar{s}_{ih} = \text{double sampling estimator of the mean of } s_i$$

The estimator z_i has been discussed by Cochran (1977). Using our notation and assuming that (i) the population size N is so large that the terms

divided by N can be ignored and (ii) the sample size n' is small with respect to the population size N so that we can use the approximation $(N-n')/(N-1) = 1$, we can write Cochran's equation (12.24) of page 333 as

$$S_{z_i z_i} = \sum (Q_h)^2 S_{s_i s_i}^{hh} / n_h + \sum Q_h (\bar{s}_{ih} - z_i)^2 / n'$$

where \sum is taken over all strata h . This formula can be extended easily and obtain a formula for the covariance of z_i and z_j as

$$S_{z_i z_j} = \sum (Q_h)^2 S_{s_i s_j}^{hh} / n_h + \sum Q_h (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / n'$$

This is an approximate formula which would suffice in most forest inventory problems. If more exact formulae are needed, the interested reader should refer to Cochran (1977).

As the covariance matrix $[S_{zz}]$ is now defined, we can write the usual variance formula

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

Case 2: The Two Samples Are Statistically Independent

Let us consider here the more general case where different, not necessarily independent volume regression functions may be used in different strata. As in the previous section, we define the giant size vector

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

and

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

$$= \begin{bmatrix} Q_1 \bar{s}_{11} & Q_1 \bar{s}_{21} & \dots & Q_1 \bar{s}_{m1} & Q_2 \bar{s}_{12} & \dots & Q_L \bar{s}_{mL} \end{bmatrix}$$

The giant size covariance matrix $[S_{BB}]$ of $[B]$ has been defined in the previous section. Because Q_h is no longer known without error, the giant covariance matrix $[S_{ZZ}]$ would, however, be different. To calculate it, we shall use the following result.

If u_1 and u_2 are two random variables that are statistically independent of two other random variables v_1 and v_2 , then the covariance of $z_1 = u_1 v_1$ and $z_2 = u_2 v_2$ is approximately equal to

$$S_{z_1 z_2} = u_1 u_2 S_{v_1 v_2} + v_1 v_2 S_{u_1 u_2}$$

Applying this rule to the variables $Q_h \bar{s}_{ih}$ and $Q_k \bar{s}_{jk}$ and using the formulae of the variances and covariances of Q_h, Q_k, \bar{s}_{ih} and \bar{s}_{jk} given in this and previous sections, we can state that

(1) the variance of $Q_h \bar{s}_{ih}$, for any i and h ,

$$(Q_h)^2 S_{s_i s_i}^{hh} / n_h + (\bar{s}_{ih})^2 Q_h (1-Q_h) / n'$$

(2) the covariance of $Q_h \bar{s}_{ih}$ and $Q_h \bar{s}_{jh}$ for given h and $i \neq j = 1, 2, \dots, m$ is

$$(Q_h)^2 S_{s_i s_j}^{hh} / n_h + \bar{s}_{ih} \bar{s}_{jh} Q_h (1-Q_h) / n'$$

(3) the covariance of $Q_h \bar{s}_{ih}$ and $Q_k \bar{s}_{jk}$ for $h \neq k = 1, 2, \dots, L$ and $i, j = 1, 2, \dots, m$, is

$$(-\bar{s}_{ih} \bar{s}_{jk} Q_h Q_k / n')$$

since the term $Q_h Q_k S_{s_i s_j}^{hk} / n_h = 0$ due to the statistical independence between the plots of different strata.

To calculate $[S_{ZZ}]$ it is, however, much more convenient to use matrix operations. We have already defined the matrix $[S_{SS}^{hh}]$ and we shall now define the product of the mean vectors

$$[\bar{s}^h][\bar{s}^k]' = \begin{bmatrix} \bar{s}_{1h} \bar{s}_{1k} & \bar{s}_{1h} \bar{s}_{2k} & \dots & \bar{s}_{1h} \bar{s}_{mk} \\ \bar{s}_{2h} \bar{s}_{1k} & \bar{s}_{2h} \bar{s}_{2k} & \dots & \bar{s}_{2h} \bar{s}_{mk} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{s}_{mh} \bar{s}_{1k} & \bar{s}_{mh} \bar{s}_{2k} & \dots & \bar{s}_{mh} \bar{s}_{mk} \end{bmatrix}$$

Then, for a given h ,

$$[S_{ZZ}^{hh}] = (Q_h)^2 / n_h [S_{SS}^{hh}] + (Q_h(1-Q_h) / n') [\bar{s}^h][\bar{s}^h]'$$

and for given $h \neq k$,

$$[S_{ZZ}^{hk}] = -(Q_h Q_k / n') [\bar{s}^h][\bar{s}^k]' = [S_{ZZ}^{kh}]'$$

Consequently, the covariance matrix of $[Z]$ can be written as

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

and finally

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

= double sampling estimator of μ , the mean volume per unit area

and

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

= estimator of the variance of w .

Example 2 - Assume that the stratum sizes of 17164, 19056 and 6116 acres of Example 1 were not known without error; they were estimated from a set of photo-points randomly located on aerial photographs of the forest and analyzed as to the stratum they happen to fall. More specifically, let us assume that $n' = 1253$ randomly selected photo-points were classified by stratum and $n'_1 = 508$, $n'_2 = 564$ and $n'_3 = 181$ points were found to fall within stratum 1, 2, and 3 respectively. Then, we calculate first the following statistics.

(1) The relative size of the three strata are estimated as

$$Q_1 = n'_1 / n' = 508 / 1253 = .40542698$$

$$Q_2 = n'_2 / n' = 564 / 1253 = .45011971$$

$$Q_3 = n'_3 / n' = 181 / 1253 = .14445331$$

(2) As the total forest area is known to be equal to $A = 42336$ acres, the area of each stratum A_h is estimated as

$$A_1 = An'_1 / n' = 17164 \text{ acres}$$

$$A_2 = An'_2 / n' = 19056 \text{ acres}$$

$$A_3 = An'_3 / n' = 6116 \text{ acres}$$

These were the areas assumed known without error in Example 1

(3) The variances and covariances of Q_h and Q_k are estimated as

$$S_{QQ}^{11} = Q_1(1-Q_1) / n' = .00019238304$$

$$S_{QQ}^{12} = -Q_1 Q_2 / n' = -.00014564300$$

$$S_{QQ}^{22} = Q_2(1-Q_2) / n' = .00019753548$$

$$S_{QQ}^{13} = -Q_1 Q_3 / n' = -.000046740039$$

$$S_{QQ}^{23} = -Q_2 Q_3 / n' = -.000051892485$$

$$S_{QQ}^{33} = Q_3(1-Q_3) / n' = .000098632524$$

(4) The giant size vector $[B]$ and covariance matrix $[S_{BB}]$, both of order 9 are defined in terms of the subvector $[b]$ and submatrix $[S_{bb}]$ shown in Table 1 as

$$[B] = \begin{bmatrix} [b] \\ [b] \\ [b] \end{bmatrix} \text{ and } [S_{BB}] = \begin{bmatrix} [S_{bb}] & [S_{bb}] & [S_{bb}] \\ [S_{bb}] & [S_{bb}] & [S_{bb}] \\ [S_{bb}] & [S_{bb}] & [S_{bb}] \end{bmatrix}$$

(5) The giant size vector $[Z]$ is defined as

$$[Z] = \begin{bmatrix} [z^1] \\ [z^2] \\ [z^3] \end{bmatrix} = \begin{bmatrix} Q_1 [z^1] \\ Q_2 [z^2] \\ Q_3 [z^3] \end{bmatrix} = \begin{bmatrix} Q_1 [\bar{s}^1] \\ Q_2 [\bar{s}^2] \\ Q_3 [\bar{s}^3] \end{bmatrix}$$

and the giant size covariance matrix $[S_{ZZ}]$ is defined as

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

where the formulae of the submatrices, for given h , are

$$[S_{ZZ}^{hh}] = (Q_h)^2 / n_h [S_{SS}^{hh}] + (Q_h(1-Q_h) / n') [\bar{s}^h][\bar{s}^h]'$$

and for $h \neq k$

$$[S_{ZZ}^{hk}] = -(Q_h Q_k / n') [\bar{s}^h][\bar{s}^k]' = [S_{ZZ}^{kh}]'$$

The numerical values of $[s^h]$ and $[S_{SS}^{hh}]$ are those given in Table 1 and the numerical values of $[Z^h]$ and $[S_{ZZ}^{hk}]$ are shown in Table 2.

Consequently we can now write

$$w = [B]'[Z] = 98315.308$$

= estimate of the mean biomass per acre

and

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

$$= 6525139.8 + 9836818.8 = 16361958.6$$

= estimate of the variance of w.

As the reader can verify, some of these results are close, but not identical to those obtained in Example 1. The two estimates of μ are 98316.722 and 98315.308 and the two variance components associated with the error of the biomass regression function are 9837039.2 and 9836818.8. The differences are due to round-off error; in Example 1 the relative sizes of strata were calculated from the estimated stratum sizes to the nearest acre. On the other hand the estimate of the variance component due to sample plots is somewhat larger in Example 2; 6525139.8 compared to the value 5919942.4 of Example 1. This is due to the fact that in Example 1 the effect of the error of Q_1 , Q_2 and Q_3 was ignored.

Table 2 - The numerical values of the subvectors $[Z^h]$ and submatrices $[S_{ZZ}^{hk}]$ of Example 2

$$[Z^1] = \begin{bmatrix} 57.649738 \\ 301.42358 \\ 1740.6166 \end{bmatrix}, \quad [Z^2] = \begin{bmatrix} 124.94841 \\ 741.43076 \\ 5305.2802 \end{bmatrix},$$

$$[Z^3] = \begin{bmatrix} 36.183793 \\ 285.29283 \\ 3059.5142 \end{bmatrix}$$

$$[S_{ZZ}^{11}] = \begin{bmatrix} 31.458403 & 159.36844 & 888.69397 \\ 159.36844 & 830.00163 & 4814.6545 \\ 888.69397 & 4814.6545 & 29596.314 \end{bmatrix}$$

$$[S_{ZZ}^{12}] = \begin{bmatrix} -5.7487974 & -34.112761 & -244.09259 \\ -30.057779 & -178.35971 & -1276.2463 \\ -173.57325 & -1029.9654 & -7369.8794 \end{bmatrix} = [S_{ZZ}^{21}]'$$

$$[S_{ZZ}^{22}] = \begin{bmatrix} 47.686234 & 230.60834 & 1069.7501 \\ 230.60834 & 1212.8417 & 6755.1404 \\ 1069.7501 & 6755.1404 & 51977.059 \end{bmatrix}$$

$$[S_{ZZ}^{13}] = \begin{bmatrix} -1.6647935 & -13.126143 & -140.76631 \\ -8.7044282 & -68.630476 & -736.00139 \\ -50.265053 & -396.31718 & -4250.1526 \end{bmatrix} = [S_{ZZ}^{31}]'$$

$$[S_{ZZ}^{23}] = \begin{bmatrix} -3.6082262 & -28.449230 & -305.09292 \\ -21.410836 & -168.81475 & -1810.3894 \\ -153.20444 & -1207.9476 & -12954.174 \end{bmatrix} = [S_{ZZ}^{32}]'$$

$$[S_{ZZ}^{33}] = \begin{bmatrix} 22.982543 & 125.46803 & 693.13551 \\ 125.46803 & 707.59481 & 4330.4116 \\ 693.13551 & 4330.4116 & 35393.175 \end{bmatrix}$$

Acknowledgements

This paper is based on research funded by the Research Foundation of the State University

of New York, the United States Department of Agriculture Forest Service and the Department of Energy, Grant No. 23-524.

Literature Cited

- Cochran, W. G. Sampling Techniques, 3rd Ed. John Wiley and Sons, New York, NY; 1977.
- Cunia, T. Dummy variables and some of their uses in regression analysis. In: Proceedings of the June 1973 meeting of IUFRO Subject Group S4.02, Nancy-France, Vol. 1, T. Cunia, K. Kusela and A. J. Nash (Eds), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1973.
- Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the workshop on "Tree biomass functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "Tree biomass functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.
- Cunia, T. On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986c.
- Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the workshop on "Tree biomass functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

ON THE ERROR OF FOREST INVENTORY ESTIMATES: TWO-STAGE SAMPLING OF PLOTS⁽¹⁾

Tiberius Cunia

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

The error of forest biomass estimates has two main sources, the sample plots where the trees are measured for diameter only and the regression functions that are used to estimate the biomass. Given that the biomass regressions are linear with known error, an approach is suggested to combine this error with the error of the sample plots where the sample plots are selected by a two-stage cluster sampling design.

Introduction

It is common to have forest inventory designs consisting of trees selected in two phases. In the first phase the trees of a sample of plots (or Bitterlich relascope points) are measured for diameter, species and possibly other attributes but not measured for biomass. In the second phase a sample of trees is measured for biomass in addition to diameter, species and possibly other attributes. The trees of the second phase are then used to estimate the regression function of tree biomass on diameter (and possibly other attributes), which applied to trees of the first phase yields estimates of the average biomass per acre. The error of the estimates has two main components, one component associated with each of the two sampling phases above. However, it is common to ignore the second error component (due to the biomass regression function) when the error of the average biomass per acre estimate is calculated.

An approach proposed by Cunia (1965, 1986a) can be used to combine the error from the first phase sample plots with the error from the second phase sample trees. This approach requires that the estimators be of the form

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

where (i) [b] is the second phase sample estimator of the vector of coefficients of the regression of biomass on x_1, x_2, \dots, x_m (functions of various tree attributes other than biomass) assumed to be of the linear form

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

and (ii) [z] is a vector of statistics calculated from the data of the first phase sample plots. The two vectors [b] and [z] are assumed to be statistically independent and the variance of w can be estimated by the approximate formula

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

Note that in this formula $[S_{bb}]$ denotes the covariance matrix of [b] and that the right hand side of the equation has two terms; the first term is the variance component associated with the error of the sample plots (of the first phase) and the second term is the variance component associated with the sample trees, or biomass regression (of the second phase).

The definition of the vector [z] depends on (i) the sampling design of the first phase, (ii) the parameter μ one wishes to estimate and (iii) the type of variables x used in the biomass regression function. In his paper, Cunia (1986a) assumes that (i) the sample plots of the first phase are selected by simple random sampling and (ii) the parameter μ is the mean biomass per acre. Given a biomass regression function of the form

$$\hat{y} = b_1 x_1 + b_2 x_2 + \dots + b_m x_m = [b]' [x]$$

the statistics z_1, z_2, \dots, z_m are then defined as the averages of the sample plot variables $s_i = (\sum x_i)/a, i = 1, 2, \dots, m$, where a is the plot area (in acres) and \sum is taken over the trees of a given sample plot. Note that s_i is the sum of the variables x_i of the trees from within a given plot, expressed on a per acre basis. For example, if $x_1 = 1$ and $x_2 = d$ (diameter), then

$s_1 =$ number of trees per acre of a given plot, and $s_2 =$ sum of diameters per acre of a given plot

If, for n = number of sample plots, the sample means, variances and covariances of the plot variables s_i are denoted by

$$\bar{s}_i = (\sum s_i)/n, \text{ and}$$

$$S_{s_i s_j} = \sum (s_i - \bar{s}_i)(s_j - \bar{s}_j)/(n-1)$$

where $i, j = 1, 2, \dots, m$ and \sum is taken over the n sample plots, then

$$z_i = \bar{s}_i \text{ and } S_{z_i z_j} = S_{s_i s_j} / n$$

Assuming now that [b] and $[S_{bb}]$ are given, the estimates w of μ and S_{ww} of the variance of w are given by the formulae above.

The objective of the present paper is to define the vector [z] of statistics when (i) the sample plots are selected by a two-stage cluster sampling design and (ii) the parameter μ of interest is the average biomass per acre. We shall also show how to calculate $[S_{zz}]$ an estimator of the covariance matrix of [z]. We shall assume that the biomass regression function is of the linear form and that we are given the vector [b] of re-

(1) Paper based on a set of lecture notes "On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

gression coefficients and the covariance matrix $[S_{bb}]$ of $[b]$. For a description of the methodology to calculate $[b]$ and $[S_{bb}]$ the reader is referred to Cunia (1986b, c) among others.

We shall also assume that the reader is familiar with the two-stage cluster sampling theory as described in standard textbooks as, for example, that by Cochran (1977). If not familiar with the approach to combine the error from sample plots and biomass regressions, the reader is strongly advised to read the Cunia (1986a) paper. In particular, he should understand well the definition of the plot variables s_1, s_2, \dots, s_m for both sample plots of fixed area and relascope sample points. To facilitate the discussion we shall often use sample plots to denote both cases, of plots and points.

Two-stage Cluster Sampling Applied to Forest Inventory

The clusters of trees previously defined as sample plots of fixed area (or Bitterlich sample points) can be grouped into clusters of higher order (clusters of plots) and the sampling can be done in two stages. Let us denote the clusters of plots as primary sampling units and the plots themselves as secondary units. Within this context, the trees constitute the tertiary units. For example, a large forest area (an entire country or one of its political or administrative subdivisions) can be divided first into blocks, divisions or homogeneous stands (the primary units) and each block, division or stand can be further subdivided into a certain number of sample plots (the secondary units). The trees of the original forest population are the tertiary sampling units.

The general two-stage sampling design consists of (i) a first stage where a sample of primary units is selected by some random procedure and (ii) a second stage where a separate subsample of secondary units is selected at random from each of the primary units selected in the first stage. When the units of the original population (the tertiary units) found in all the sample units of the second stage are measured for the variables of interest we have the method of two-stage sampling. But when each secondary unit of the sample of the second stage is subsampled for the tertiary units we have a three-stage sampling design.

Note that strictly speaking, two-stage sampling refers to the selection of the secondary units which are completely measured for the tertiary units found within them. The estimates are calculated for the parameters of the population of secondary units. When the estimates refer to the population of tertiary units one may wish to denote the sampling method as three-stage. It is only a matter of terminology.

To introduce the necessary terminology, notation and formulae, let us assume that (i) the forest area is divided into M homogeneous stands, (the primary units), (ii) each stand h has a known area of A_h acres, $h = 1, 2, \dots, M$ and can be subdivided either into a finite number N_h of non-overlapping

plots of fixed area "a" acres (the secondary units) or into infinitely many overlapping plots of fixed area or Bitterlich sample points of given basal area factor "c", (iii) each sample plot or point k of the stand h contains n_{hk} trees (the tertiary units), $k = 1, 2, \dots, N_h$ (or infinity) and (iv) the diameter of the g -th tree in the k -th plot (or point) of the h -th stand is denoted by d_{hkg} . Then, the two-stage random sampling design considered here can be described as follows.

In the first stage, m stands are selected by simple random sampling without replacement. Each selected stand of known area of A_h acres is subdivided (at least conceptually) into N_h non-overlapping sample plots of "a" acres each, or infinitely many overlapping plots of fixed area or Bitterlich sample points. In the second stage, n_h sample plots or points are selected from each sample stand $h = 1, 2, \dots, m$ by simple random sampling without replacement (or an equivalent systematic sampling design). All n_{hk} trees of the selected hk -th plot or point, $k = 1, 2, \dots, n_h$ are measured for their diameters d_{hkg} , $g = 1, 2, \dots, n_{hk}$.

To introduce the necessary formulae, let us assume that the k -th plot or point of the h -th sample stand is measured for its value y_{hk} = biomass per acre. Then, two-stage cluster sampling theory tells us that, for sufficiently large samples, a slightly biased, but efficient estimator of μ , the average biomass per acre for the entire forest area, is the ratio estimator

$$w = \bar{y}_R = (\sum N_h \bar{y}_h) / (\sum N_h)$$

where \bar{y}_h is the estimator of the average biomass per acre of stand h , say μ_h , calculated by the formula

$$\bar{y}_h = (y_{h1} + y_{h2} + \dots + y_{hn_h}) / n_h$$

and Σ means summation over the sample stands $h = 1, 2, \dots, m$. When N is equal to infinity, the case of the overlapping sample plots of fixed area or Bitterlich sample points, the formula becomes

$$w = \bar{y}_R = (\sum A_h \bar{y}_h) / (\sum A_h)$$

The variance of \bar{y}_R can be estimated by the formula

$$S_{\bar{y}_R \bar{y}_R} = \left(\frac{M-m}{M} \right) \left(\frac{m}{m-1} \right) \left(\sum N_h^2 (\bar{y}_h - \bar{y}_R)^2 / (\sum N_h)^2 \right) + \left(\frac{m}{M} \right) \sum \left(\frac{N_h - n_h}{N_h} \right) \left(\frac{\sum_{yy}^{shh}}{n_h} \right) / (\sum N_h)^2$$

where Σ means again summation over $h = 1, 2, \dots, m$ sample stands and s_{yy}^{shh} is an estimate of the variance within the sample stand h , that is the sample variance of the n_h plot values y_{hk} , $k = 1, 2, \dots, n_h$ and $h = 1, 2, \dots, m$. When M is large with respect to m , one can make $(M-m)/M = 1$. Also, when m is sufficiently large, one can make $m/(m-1) = 1$. If, for a given h , N_h is large relative to n_h the factor $(N_h - n_h)/N_h$ can be made

equal to 1. Finally, when N_h is infinitely large, as the case may be with overlapping sample plots of fixed area or relascope sample points, one can substitute A_h for N_h . If all four conditions above are satisfied, one can use the following approximate but simpler formula

$$s_{\bar{y}_R \bar{y}_R} = \left(\Sigma A_h^2 (\bar{y}_h - \bar{y}_R)^2 + m \Sigma A_h^2 S_{hh} / M n_h \right) / (\Sigma A_h)^2$$

Sometimes the variation within the stands as measured by S_{hh} is relatively small compared to the variance between stand means as represented by the squared differences $(\bar{y}_h - \bar{y}_R)^2$. Then, the effect of the second variance component

$$m \Sigma A_h^2 S_{hh} / M n_h (\Sigma A_h)^2$$

can be ignored and the variance of \bar{y}_R can be estimated by the formula

$$s_{\bar{y}_R \bar{y}_R} = \Sigma A_h^2 (\bar{y}_h - \bar{y}_R)^2 / (\Sigma A_h)^2$$

All the above formulae of the variance of \bar{y}_R ignore the error of the biomass regression function used for the calculation of the plot values y_{hk} . We shall now show how to include the error of the regression function into the error of the estimate \bar{y}_R of μ . We shall consider first the case of a single biomass regression applied to all trees of all stands and then extend the results to the case of different regressions applied to different stands. To facilitate our discussion, we shall assume that the biomass regression function is of the form

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

where d is the tree diameter and the definition of $[x]$ is obvious.

Case 1 - Same biomass regression function applied to all stands

The biomass per acre of the hk -th plot or point can be defined by the usual formula as

$$y_{hk} = b_1 s_{1hk} + b_2 s_{2hk} + b_3 s_{3hk}$$

where

- s_{1hk} = number of trees per acre of the hk -th plot or point
- s_{2hk} = average sum of tree diameters per acre of the hk -th plot or point
- s_{3hk} = average sum of squared diameters per acre of the hk -th plot or point

If we write

$$\bar{y}_h = b_1 \bar{s}_{1h} + b_2 \bar{s}_{2h} + b_3 \bar{s}_{3h}$$

where \bar{s}_{1h} , \bar{s}_{2h} , and \bar{s}_{3h} are the averages of the n_{hk} sample values s_{1hk} , s_{2hk} , and s_{3hk} respectively of the sample stand h , we can write the estimate of μ , the average biomass per acre, as

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

where, for Σ meaning summation over the m sample stands h ,

$$z_1 = \Sigma A_h \bar{s}_{1h} / \Sigma A_h = \Sigma N_h \bar{s}_{1h} / \Sigma N_h$$

$$z_2 = \Sigma A_h \bar{s}_{2h} / \Sigma A_h = \Sigma N_h \bar{s}_{2h} / \Sigma N_h$$

$$z_3 = \Sigma A_h \bar{s}_{3h} / \Sigma A_h = \Sigma N_h \bar{s}_{3h} / \Sigma N_h$$

Each z_i value, $i = 1, 2, 3$, is a ratio estimator of the type \bar{y}_R defined above. Its variance formula has been given above. Extending this formula, we can write the covariance of z_i and z_j as

$$s_{z_i z_j} = \left(\frac{M-m}{M} \right) \left(\frac{m}{m-1} \right) \Sigma N_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (\Sigma N_h)^2 + \left(\frac{m}{M} \right) \Sigma \left(\frac{N_h - n_h}{N_h} \right) \left(\frac{N_h^2 S_{ij}^{hh}}{n_h} \right) / (\Sigma N_h)^2$$

where Σ is taken over the m sample stands and S_{ij}^{hh} is the sample covariance of the n_h pairs of values s_{ihk} and s_{jkh} of the sample stand h , that is

$$S_{ij}^{hh} = \Sigma (s_{ihk} - \bar{s}_{ih}) (s_{jkh} - \bar{s}_{jh}) / (n_h - 1)$$

where now Σ means summation within the stand h over plot $k = 1, 2, \dots, n_h$. When we can approximate the values $(M-m)/M$, $m/(m-1)$ and $(N_h - n_h)/N_h$ by 1, and if N_h is infinitely large for all $h = 1, 2, \dots, m$, the formula above simplifies to

$$s_{z_i z_j} = \left(\Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) + m \Sigma A_h^2 S_{ij}^{hh} / M n_h \right) / (\Sigma A_h)^2$$

Finally, when the covariance within clusters is small relative to the covariance between cluster means, one can use the formula

$$s_{z_i z_j} = \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (\Sigma A_h)^2$$

Case 2 - Different biomass regression functions applied to different stands

Many times, the various stands are defined in terms of forest type or site for which the trees are of a different shape, and thus, different regression functions have been constructed for different stands. Then, let us assume that

$$[b]^h = [b_{1h} \ b_{2h} \ b_{3h}]$$

is the regression applied to stand h , that the covariance matrix of $[b]^h$ is as usually denoted as $[S_{bb}^h]$, and that the m vectors of regression coefficients $[b]^h$ have been calculated by a least squares technique (or some of its extensions) such that the covariance matrix of $[b]^h$ with $[b]^k$ denoted here as $[S_{bb}^{hk}]$ can be calculated. If we define the giant size vector $[B]$ as

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

then, the covariance matrix of $[B]$ can be written as

$$[S_{BB}] = \begin{bmatrix} [s_{bb}^{11}] & [s_{bb}^{12}] & \dots & [s_{bb}^{1m}] \\ [s_{bb}^{12}] & [s_{bb}^{22}] & \dots & [s_{bb}^{2m}] \\ \vdots & \vdots & \ddots & \vdots \\ [s_{bb}^{1m}] & [s_{bb}^{2m}] & \dots & [s_{bb}^{mm}] \end{bmatrix}$$

For the special case where the same regression function is applied to stands h and k, then

$$[S_{bb}^{hh}] = [S_{bb}^{kk}] = [S_{bb}^{hk}]$$

while for the case of statistically independent vectors $[b^h]$ and $[b^k]$, we have $[S_{bb}^{hk}] = [0]$.

The biomass per acre of the hk-th plot or point can be defined as

$$y_{hk} = b_{1h} s_{1hk} + b_{2h} s_{2hk} + b_{3h} s_{3hk}$$

where s_{1hk} , s_{2hk} , and s_{3hk} are defined as in Case 1 above. Also, the average biomass per acre of stand h is estimated by

$$\bar{y}_h = b_{1h} \bar{s}_{1h} + b_{2h} \bar{s}_{2h} + b_{3h} \bar{s}_{3h} = [b^h]' [s^h]$$

where \bar{s}_{1h} , \bar{s}_{2h} , and \bar{s}_{3h} are again defined as in Case 1 above. Then, the two-stage sampling estimator of μ can be written as

$$\begin{aligned} w &= (A_1 \bar{y}_1 + A_2 \bar{y}_2 + \dots + A_m \bar{y}_m) / (A_1 + A_2 + \dots + A_m) \\ &= (A_1 / \Sigma A_k) (b_{11} \bar{s}_{11} + b_{21} \bar{s}_{21} + b_{31} \bar{s}_{31}) \\ &+ (A_2 / \Sigma A_k) (b_{12} \bar{s}_{12} + b_{22} \bar{s}_{22} + b_{32} \bar{s}_{32}) \\ &+ \dots \\ &+ (A_m / \Sigma A_k) (b_{1m} \bar{s}_{1m} + b_{2m} \bar{s}_{2m} + b_{3m} \bar{s}_{3m}) \\ &= [B]' [Z] \end{aligned}$$

where

$$\begin{aligned} [Z]' &= [A_1 [\bar{s}_1]' \quad A_2 [\bar{s}_2]' \quad \dots \quad A_m [s_m]' / \Sigma A_k \\ &= [[z^1]' \quad [z^2]' \quad \dots \quad [z^m]'] \end{aligned}$$

with $\Sigma A_k = A_1 + A_2 + \dots + A_m$, and

$$[z^h]' = (A_h / \Sigma A_k) [\bar{s}_{1h} \quad \bar{s}_{2h} \quad \bar{s}_{3h}]$$

The covariance matrix of $[z^h]$ is calculated as in Case 1 above and denoted as $[S_{ZZ}^{hh}]$. As the sample plots or points of stand h are selected independently of the sample plots or points of stand k, the covariance matrix of $[z^h]$ with $[z^k]$ is equal to zero; that is $[S_{ZZ}^{hk}] = [0]$ when $h \neq k$. Consequently

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

An Illustrative Example

To better see how the formulae of the previous section apply to an actual set of sample data, let us consider the following numerical example.

Example - Assume that a forest area of $A = 42336$ acres is divided into 140 blocks of sizes varying from about 100 to about 500 acres each. We are given that (i) $m = 16$ blocks are selected by simple random sampling without replacement from the $M = 140$ population blocks, (ii) each block so selected is measured for its area, say A_h in acres, (iii) n_h one-tenth acre sample plots are selected by simple random sampling without replacement from the h-th sample block, with the sample size n_h being approximately proportional to A_h , and (iv) the total sample size is $n = 235$. It is implicitly assumed that the blocks and plots are non-overlapping. The values A_h and n_h are shown by block in Table 1.

To calculate the biomass we shall use the regression function already used by Cunia (1986a). This regression is assumed to be of the linear (parabolic) form

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

where y = tree biomass and d = tree diameter, with obvious definitions for x_1, x_2 and x_3 . The estimates $[\hat{b}]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ are shown in the Cunia (1986a) paper, and, for convenience they are also given below

$$[b]' = [5.1818118 \quad -25.653078 \quad 12.988357]$$

and

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

This regression function implies that the plot variables are defined as

$$\begin{aligned} s_1 &= \Sigma x_1 / .10 = 10 \Sigma (1) \\ &= \text{number of trees per acre} \\ s_2 &= \Sigma x_2 / .10 = 10 \Sigma d \\ &= \text{sum of diameters per acre, and} \\ s_3 &= \Sigma x_3 / .10 = 10 \Sigma d^2 \\ &= \text{sum of squared diameters per acre,} \end{aligned}$$

where Σ means summation over the trees of a given plot.

Consider now the sample of plots. The values s_1, s_2 and s_3 are calculated for each plot separately. The individual plot values are listed by Cunia (1986d, Tables 1 and 2) and are not repeated here. For the purpose of this example, the 235 plots are distributed arbitrarily among the 16 blocks, according to the number n_h of plots that they should contain. More specifically, the first $n_1 = 13$ plots were assigned to block 1, the next $n_2 = 15$ plots were assigned to block 2, etc. with the last $n_{16} = 18$ plots assigned to the last block 16.

Using the terminology and notation of the previous section, the secondary units are the

Table 1 - The area A_h (acres), the number n_h of sample plots, the averages of s_1 , s_2 (inches), s_3 (squared inches) and y (pounds) and the variance V of y within blocks by block number h

h	A_h	n_h	\bar{s}_1	\bar{s}_2	\bar{s}_3	\bar{y}	V
1	280	13	210.76923	1243.5154	8765.4079	83040.42	4949333039
2	302	15	200.66667	1175.9533	8444.4321	80552.29	3840381639
3	452	22	245.45455	1403.3682	9800.8083	92567.58	5702946485
4	216	10	149.00000	1010.4200	8361.1126	83448.82	6457081439
5	384	18	205.55556	1323.7056	11052.1138	110656.83	6406414888
6	290	14	173.57143	1380.9929	15067.3455	161172.76	4858967459
7	242	12	279.16667	1588.3583	11129.5609	105255.02	3023986664
8	315	15	230.66667	1448.6333	12156.5293	121926.71	7706768567
9	344	17	232.35294	1289.7118	8270.5904	75540.31	2658394967
10	287	14	166.42857	958.6500	6639.4076	62505.07	4666270466
11	472	23	226.52174	1320.5826	9660.5531	92771.50	3502826347
12	130	6	218.33333	1552.1833	13833.9085	140992.82	7053833420
13	212	10	264.00000	1801.9300	16101.5319	164275.39	8978786810
14	274	13	251.53846	1753.6385	15748.0664	160858.71	3564467131
15	319	15	271.33333	1504.5600	9355.6103	84323.41	2743051584
16	382	18	258.88889	1626.2444	12780.5912	125622.22	7090217424

sample plots and the primary units are the blocks. There is a total of $N = 423360$ possible one-tenth acre plots and $M = 140$ blocks. The first stage sample size is $m = 16$ sample blocks and the second stage sample size is $n = 235$ sample plots. The total area of the 16 sample blocks is

$$\Sigma A_h = A_1 + A_2 + \dots + A_{16} = 280 + 302 + \dots + 382 = 4901 \text{ acres}$$

The first step is the calculation of the averages of the variables s_1 = number of trees per acre, s_2 = sum of tree diameters per acre, and s_3 = sum of squared tree diameters per acre for each sample block $h = 1, 2, \dots, 16$ separately. The 16 sets of values \bar{s}_{1h} , \bar{s}_{2h} , and \bar{s}_{3h} are listed in Table 1.

We continue with the calculation of the vector $[z]$ defined as

$$[z] = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 225.58049 \\ 1390.1177 \\ 10830.566 \end{bmatrix}$$

where, for Σ meaning summation over the sample blocks $h = 1, 2, \dots, 16$

$$z_1 = \Sigma A_h \bar{s}_{1h} / \Sigma A_h = 110570 / 4901 = 225.58049$$

= two stage sampling estimate of the average number of trees per acre of the entire forest area

$$z_2 = \Sigma A_h \bar{s}_{2h} / \Sigma A_h = 6812966.8 / 4901 = 1390.1177$$

= two stage sampling estimate of the average sum of tree diameters per acre of the entire forest area, and

$$z_3 = \Sigma A_h \bar{s}_{3h} / \Sigma A_h = 53080603 / 4901 = 10830.566$$

= two stage sampling estimate of the average sum of squared tree diameters per acre of the entire forest area

The ij -th element of the covariance matrix $[S_{zz}]$ of $[z]$ can be calculated by the formula, say Main Formula

$$S_{z_i z_j} = \left(\frac{M-m}{M} \right) \left(\frac{m}{m-1} \right) \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (\Sigma A_h)^2 + \left(\frac{m}{M} \right) \Sigma \left(\frac{10A_h - n_h}{10A_h} \right) \left(\frac{A_{hh}^{ij}}{n_h} \right) / (\Sigma A_h)^2$$

where Σ is taken over $h = 1, 2, \dots, 16$ sample clusters, and $N_h = 10A_h$. This requires the calculation of 16 covariance matrices of the variables s_1 , s_2 , and s_3 within blocks, say $[S_{hh}^{ij}]$, with its ij -th elements defined by the formula

$$\Sigma (s_{ihg} - \bar{s}_{ih}) (s_{jhg} - \bar{s}_{jh}) / (n_h - 1)$$

for $i, j = 1, 2, 3$, where Σ is taken over the n_h sample values s_{ihg} and s_{jhg} of sample plots $g = 1, 2, \dots, n_h$ within sample block h . The ij -th element of the covariance matrix of $[z]$ can also be calculated by any of the following alternative formulae.

Short-cut Formula 1

$$S_{z_i z_j} = m \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (m-1) (\Sigma A_h)^2 + m \Sigma A_h^2 S_{hh}^{ij} / n_h M (\Sigma A_h)^2$$

Short-cut Formula 2

$$S_{z_i z_j} = m \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (m-1) (\Sigma A_h)^2, \text{ and}$$

Short-cut Formula 3

$$S_{z_i z_j} = \Sigma A_h^2 (\bar{s}_{ih} - z_i) (\bar{s}_{jh} - z_j) / (\Sigma A_h)^2$$

It is more convenient, however, to define the covariance matrix $[S_{zz}]$ directly by matrix operations. For this we define first the 16 by 3 matrix of weighted differences

$$[V] = \begin{bmatrix} A_1(\bar{s}_{11}-z_1) & A_1(\bar{s}_{21}-z_2) & A_1(s_{31}-z_3) \\ A_2(\bar{s}_{12}-z_1) & A_2(\bar{s}_{22}-z_2) & A_2(\bar{s}_{32}-z_3) \\ \vdots & \vdots & \vdots \\ A_{16}(\bar{s}_{1,16}-z_1) & A_{16}(\bar{s}_{2,16}-z_2) & A_{16}(\bar{s}_{3,16}-z_3) \end{bmatrix}$$

$$= \begin{bmatrix} -4147.1529 & -41048.644 & -578244.24 \\ -7523.9750 & -64677.634 & -720612.42 \\ \vdots & \vdots & \vdots \\ 12723.808 & 90200.423 & 744909.67 \end{bmatrix}$$

and then, the weighted sum of 16 matrices $[S_{ss}^{hh}]$, say

$$[S_{ss}^{hh}] = \Sigma \left(\frac{10A_{hh}-n_h}{10A_h} \right) \left(\frac{A_h^2}{n_h} \right) [S_{ss}^{hh}]$$

$$= \begin{bmatrix} 2012834975 & 10315563210 & 54028872200 \\ 10315563210 & 60558424930 & 424677708600 \\ 54028872200 & 424677708600 & 4484694027000 \end{bmatrix}$$

or the corresponding matrix for the short-cut formulae

$$[S_{ss}^*] = \Sigma (A_h^2/n_h) [S_{ss}^{hh}]$$

$$= \begin{bmatrix} 2022524766 & 10365134370 & 54287700110 \\ 10365134370 & 60849065330 & 426712725100 \\ 54287700110 & 426712725100 & 4506190215000 \end{bmatrix}$$

Using the Main Formula above, the covariance matrix $[S_{zz}]$ is calculated as

$$[S_{zz}] = \left(\frac{M-n}{M} \right) \left(\frac{m}{m-1} \right) [V]'[V]/(\Sigma A_h)^2 + \left(\frac{m}{M} \right) [S_{ss}^*]/(\Sigma A_h)^2$$

$$= \begin{bmatrix} 75.342297 & 365.40810 & 1591.8295 \\ 365.40810 & 2687.0073 & 25799.155 \\ 1591.8295 & 25799.155 & 397813.92 \end{bmatrix}$$

Consequently,

$$w = [b]'[z] = 106179.37 \text{ pounds}$$

= two-stage cluster sampling estimate of μ , the average biomass per acre for the entire forest area

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

$$= 51805483 + 11063125 = 62868608$$

= estimate of the variance of w

$$\sqrt{S_{ww}} = 7928.9727$$

= estimate of the standard error of w

and using a value $t = 2$,

$$w \pm t \sqrt{S_{ww}} = (106179 \pm 15858) \text{ pounds}$$

= 95 percent confidence limits of μ

Using now Short-cut Formula 1 we obtain the values

$$[S_{zz}] = m[V]'[V]/(m-1)(\Sigma A_h)^2 + m[S_{ss}^*]/M(\Sigma A_h)^2$$

$$= \begin{bmatrix} 83.874242 & 406.46033 & 1765.2882 \\ 406.46033 & 2997.9220 & 28877.038 \\ 1765.2882 & 28877.038 & 446493.73 \end{bmatrix}$$

$$S_{ww} = 58183864 + 11063125 = 69246988$$

and

$$w \pm t \sqrt{S_{ww}} = (106179 \pm 16643) \text{ pounds}$$

for an overestimation of the confidence interval of

$$(100)(16643 - 15858)/(15858) = 4.95 \text{ percent}$$

Using Short-cut Formula 2, we obtain

$$[S_{zz}] = m[V]'[V]/(m-1)(\Sigma A_h)^2$$

$$= \begin{bmatrix} 74.251111 & 357.14324 & 1506.9885 \\ 357.14324 & 2708.4035 & 26846.747 \\ 1506.9885 & 26846.747 & 425053.37 \end{bmatrix}$$

$$S_{ww} = 55707438 + 11063125 = 66770563$$

and

$$w \pm t \sqrt{S_{ww}} = (106179 \pm 16343) \text{ pounds}$$

for an overestimation of the confidence interval of

$$(100)(16343 - 15858)/(15858) = 3.06 \text{ percent}$$

Finally, using Short-cut Formula 3, we obtain

$$[S_{zz}] = [V]'[V]/(\Sigma A_h)^2$$

$$= \begin{bmatrix} 69.610416 & 334.82178 & 1412.8017 \\ 334.82178 & 2539.1282 & 25168.826 \\ 1412.8017 & 25168.826 & 398487.54 \end{bmatrix}$$

$$S_{ww} = 52225723 + 11063125 = 63288848$$

and

$$w \pm t \sqrt{S_{ww}} = (106179 \pm 15911) \text{ pounds}$$

for an overestimation of only

$$(100)(15911 - 15858)/(15858) = .33 \text{ percent}$$

We shall now consider the common procedure of calculating (1) the biomass of each tree by the regression function, (2) the biomass of the plot hk (k -th plot of h -th sample block) by summing up the biomass of all trees within plots, (3) the biomass per acre of plot hk , say y_{hk} by dividing the plot biomass by the plot area (one-tenth of an acre), (4) the average biomass per acre for each sample block $h = 1, 2, \dots, m$, say

$\bar{y}_h = \Sigma y_{hk} / n_h$, where Σ means summation over plot $k = 1, 2, \dots, n_h$ within block h and finally, (5) the estimate w of μ and the estimate of the variance of w by the formulae

$$w = \Sigma A_h \bar{y}_h / \Sigma A_h$$

and

$$S_{ww} = \left(\frac{M-m}{M} \right) \left(\frac{m}{m-1} \right) \Sigma A_h^2 (\bar{y}_h - w)^2 / (\Sigma A_h)^2 + \left(\frac{m}{M} \right) \Sigma \left(\frac{10A_h - n_h}{10A_h} \right) \left(\frac{A_h^2 S_{yy}^{hh}}{n_h} \right) / (\Sigma A_h)^2$$

where

$$S_{yy}^{hh} = \Sigma (y_{hk} - \bar{y}_h)^2 / (n_h - 1)$$

Of course, Σ of S_{ww} means summation over $h = 1, 2, \dots, 16$ sample blocks and Σ of S_{yy}^{hh} means summation over plots $k = 1, 2, \dots, n_h$ within block h .

Because the hk -th plot values s_{1hk} , s_{2hk} , and s_{3hk} have already been calculated, there is no need to calculate the individual tree biomass. We simply calculate the biomass per acre y_{hk} of plot hk by the formula

$$y_{hk} = b_1 s_{1hk} + b_2 s_{2hk} + b_3 s_{3hk} = [b]' [s_{hk}]$$

These values are listed by Cunia (1986d, Tables 7,8) and are not given here. Taking into account the classification of these plots into the 16 blocks, we continue with the calculation of the block average biomass per acre \bar{y}_h and the sample variance within block S_{yy}^{hh} by the formulae above. These are listed in Table 1.

Finally, using these block values of means and variances, the reader can verify that

$$w = (\Sigma A_h \bar{y}_h) / (\Sigma A_h) = 520385103/4901 = 106179.37$$

and, by the formula above,

$$S_{ww} = 49340874 + 2464609 = 51805483$$

As the reader can verify, the estimate w of the average biomass per acre remains the same, while the estimate of the variance of w is equal only to the error component due to sample plots; the error component due to the biomass regression has been ignored. The variance is, thus, underestimated by

$$(100) (11063125) / (62868608) = 17.60 \text{ percent}$$

or in terms of standard errors

$$(100) (\sqrt{62868608} - \sqrt{51805483}) / (\sqrt{62868608}) = 9.22 \text{ percent}$$

Acknowledgements

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

Literature Cited

- Cochran, W.G. Sampling Techniques, 3rd. Ed. John Wiley and Sons, New York, NY; 1977.
- Cunia, T. Some theory on the reliability of volume estimates in a forest inventory sample. Forest Science, 11: 115-128; 1965.
- Cunia, T. Error of the forest inventory estimates: its main components. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.
- Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986c.
- Cunia, T. On the error of biomass estimates in forest inventories: Part2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (86-001), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

ON THE ERROR OF FOREST INVENTORY ESTIMATES:

DOUBLE SAMPLING WITH REGRESSION^{1/}

Tiberius Cunia

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

The error of biomass estimates in forest inventory contains a component due to the sample plots and another component due to the biomass regressions. A method is shown to estimate the error of the sample plots when they are selected by a double or two-phase sampling design. The error of the biomass regression function is assumed given.

Introduction

Most of the sampling designs for forest inventory consist of a relatively large sample of trees selected in clusters defined as trees growing within plots of fixed area or counted by a relascope at a Bitterlich sample point. These trees are measured for diameter, species and possibly other attributes other than biomass; and their biomass is estimated by means of previously determined regression functions. Using the estimated tree biomass one can then calculate estimates of the average biomass per unit area (by tree or stand classes). When the error of these estimates is calculated, however, the error of the biomass regressions is commonly ignored; only the error of the sample plots (or points) is taken into account.

An approach to combine the error of the biomass regression with that of the sample plots (or points) was proposed by Cunia (1965, 1986a). This approach requires that the true biomass regression function be of the form

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

where y is the tree biomass, [x] is the vector of known tree attributes other than biomass and [\beta] is the vector of unknown regression coefficients. Note that [] and []' denote vectors or matrices and their transposes respectively. The vector [\beta] is estimated from a sample of trees measured for biomass y and variables [x] and we shall use the notation [b] and [S_{bb}] to denote the estimator of [\beta] and covariance matrix of [b] respectively.

^{1/}Paper based on a set of lecture notes: "On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots". Faculty of Forestry Miscellaneous Publication Number 9 (86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

The approach also requires that the parameter μ of interest be expressed as the product of two vectors, that is,

$$\mu = [\beta]'[\mu_z] = \mu_1 \mu_{z_1} + \mu_2 \mu_{z_2} + \dots + \mu_m \mu_{z_m}$$

where μ_{z_i} is the expected value (mean) of a function of the variable x_i , $i = 1, 2, \dots, m$. If μ is the average biomass per acre, then μ_{z_i} represents the expected values of x_i expressed on a per acre basis. For example, if x_i is the tree diameter d, the μ_{z_i} is (i) the average tree diameter, when μ is defined as the mean biomass per tree or (ii) average sum of tree diameters per acre, when μ is defined as mean biomass per acre. Similarly, when $x_i = 1$, then $\mu_{z_i} = 1$ or $\mu_{z_i} =$ mean number of trees per acre depending on whether μ is the mean biomass per tree or per acre respectively.

It is assumed that the sample plots (or points) provide estimates z_i of μ_{z_i} and estimates $S_{z_i z_j}$ of the covariance of z_i and z_j , for $i, j = 1, 2, \dots, m$. Using the matrix notation, we define the vector [z], the estimate of $[\mu_z]$ and the matrix [S_{zz}], the estimate of the covariance matrix of [z]. Then, for the case where [b] and [z] are statistically independent, Cunia (1965, 1986a) has shown that

$$w = [b]'[z] = \text{estimator of } \mu$$

and

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

= estimator of the variance of w

Note how the variance of w can be viewed as having two additive components; one due to the error of the sample plots, the other due to the biomass regression.

The definition of the vector [z] depends on (i) the sampling design by which the sample plots (or points) are selected, (ii) the parameter μ one wishes to estimate and (iii) the variables x_i used in the biomass regression. In his paper, Cunia (1986a) assumes that (i) the sample plots (or points) are selected by simple random sampling (or by a systematic sampling procedure that is equivalent to simple random sampling) and (ii) the parameter μ to estimate is the average biomass per plot. If the regression function used to calculate the tree biomass is $\hat{y} = [b]'[x]$ then, the estimators z_i are statistics based on the plot (or point) variables s_i defined as $(\sum x_i)/a$ for sample plots (or $(\sum x_i/a_k)$ for sample points) where a is the plot area in acres (or the factor to convert the variable x_i of the tree k of the sample point to a per acre basis). More specifically, for n = number of sample plots (or points),

$$z_i = \bar{s}_i = \sum s_i / n$$

and

$$S_{z_i z_j} = S_{s_i s_j} / n$$

$$= \sum (s_i - \bar{s}_i)(s_j - \bar{s}_j) / (n-1)$$

The objective of this paper is to extend the Cunia (1986a) approach from the simple random

sampling method by which the plots are selected to the method of double or two-phase sampling. More specifically we shall show how to calculate estimators $[z]$ and $[S_{zz}]$ of $[\mu_z]$ and $[\sigma_{zz}]$ respectively by double sampling. We shall still be concerned with the estimator w of the mean biomass per acre μ . We shall assume that we are given the estimates $[b]$ and $[S_{bb}]$ of the biomass regression function. For some of the methodology to calculate $[b]$ and $[S_{bb}]$ the interested reader is referred to Cunia (1986b,c) among others.

We shall also assume that the reader is familiar with the theory of double or two-phase sampling as described in standard texts on sampling techniques as that by Cochran (1977), for example. Before proceeding further with the procedures described in this paper, the reader is strongly advised to become familiar with the approach of taking into account the error of biomass regressions as described by Cunia (1986a). In particular he should understand very well the definition of the plot (or point) variables s , since this definition will be assumed known when the statistics z and their variances and covariances are defined. To simplify the discussion we shall use the terminology "plots" to denote both plots of fixed area and sample points; once the variables s are calculated, it does not matter whether the sample units are plots or points.

Double Sampling with Regression Applied to Forest Inventory

Sometimes, there is an auxiliary plot variable, say v , highly correlated with the plot biomass, say u . The variable v is easy to measure, but it is of little interest; and it is linearly related to the variable of interest u , that is relatively difficult to measure. Then, to estimate μ , the mean biomass per acre, one can use, with high efficiency, a double sampling with regression estimator. For example, let us assume that good, large scale aerial photography of a given forest area of interest exists. After locating a plot on an aerial photograph, we shall also assume that, by photogrammetry and photo-interpretation of the trees of the plot, one is able to derive an estimate v of the biomass per acre, that is highly correlated with the true biomass per acre u of the same plot as measured on the ground.

Consider now the following two-phase or double sampling with regression estimator design. In the first phase, n^* photo-plots are selected by simple random sampling without replacement, their values v are determined by photogrammetry and photo-interpretation and the statistics calculated from this sample are the average \bar{v}^* and variance S_{vv}^* . In the second phase, n photo-plots are selected from the n^* plots of the first phase, again by simple random sampling without replacement. The plots are visited on the ground and measured for their true values u of biomass per acre. The sample data of the n pairs of values u and v provide the sample averages \bar{u} and \bar{v} , variances and covariance S_{uu} , S_{vv} , and S_{uv} , and the linear regression and correlation coeffi-

cients $c = S_{uv}/S_{vv}$ and $r = S_{uv}/\sqrt{S_{uu}S_{vv}}$ respectively. Then, the double sampling with regression estimator of the mean biomass per acre μ , and the estimator of its variance are defined respectively as

$$\bar{y}_r = \bar{u} - c(\bar{v} - \bar{v}^*)$$

and

$$S_{\bar{y}_r \bar{y}_r} = \left(\frac{n^* - n}{n^*} \right) S_{uu|v} \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + \left(\frac{N-n^*}{N} \right) \left(\frac{S_{uu}}{n^*} \right)$$

where N is the size of the finite population (possible number of photo-plots) and $S_{uu|v}$ is the estimator of the conditional variance of u given v calculated by the following formula, where Σ means summation over the n plots of the second phase subsample,

$$S_{uu|v} = \Sigma (u_h - u_h')^2 / (n-2) = (n-1)S_{uu}(1-r^2) / (n-2)$$

where

$$u_h' = (\bar{u} - c\bar{v}) + cv_h$$

= regression estimate of u_h of plot h when $v = v_h$

Ordinarily, the population size N is infinite in size or very large compared to the first phase sample size n^* . Then, the finite population correction factor $(N-n^*)/N$ can be made equal to 1. We shall assume here that this is always the case and, thus, unless stated otherwise, we shall make $(N-n^*)/N = 1$.

Sometimes the first phase sample size n^* may be large with respect to the second phase subsample size n and, thus, $(n^*-n)/n$ can be made equal to 1. Then, the variance formula becomes

$$S_{\bar{y}_r \bar{y}_r} = S_{uu|v} \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + \frac{S_{uu}}{n^*}$$

When n is sufficiently large, the difference between \bar{v} and \bar{v}^* may become sufficiently small so that the effect of the term $(\bar{v} - \bar{v}^*)^2 / (n-1)S_{vv}$ may become negligibly small. Then, the variance formula changes to

$$S_{\bar{y}_r \bar{y}_r} = (n^*-n)S_{uu|v}/nn^* + S_{uu}/n^* = S_{uu|v}/n - S_{uu|v}/n^* + S_{uu}/n^* = S_{uu|v}/n + (S_{uu} - S_{uu|v})/n^*$$

Note that in this last formula we have not made $(n^*-n)/n = 1$. Because, for the least squares linear regression, we have

$$S_{uu|v} = (n-1)(1-r^2)S_{uu}/(n-2) \approx (1-r^2)S_{uu}$$

we can further write

$$S_{\bar{y}_r \bar{y}_r} = S_{uu|v}/n + r^2 S_{uu}/n^*$$

Finally, when both n and n^* are sufficiently large and such that $(n^*-n)/n^*$ can be made equal to 1 and the effect of the factors $(\bar{v} - \bar{v}^*)^2 / (n-1)S_{vv}$ and $1/n^*$ is negligibly small, we can write

$$S_{\bar{y}_r \bar{y}_r} \approx S_{uu|v}/n + S_{uu}/n^* \approx S_{uu|v}/n$$

Seldom if ever one can determine the true value u of the ground plots. The usual procedure is to use a biomass regression function, estimate the biomass y of each tree in the plot by the regression value \hat{y} and then calculate an estimate \hat{u} of the true biomass u of the plot by adding the values \hat{y} of all the trees in that plot. By using \hat{u} instead of u in the formulae above results in an additional source of error (due to the biomass regression function) which is not accounted for when the variance of \bar{y}_r is calculated. To take it into account, we shall use the following procedure.

We start with the assumption that u , the true value, is not known but can be estimated by the regression value

$$\hat{u} = b_1 s_1 + b_2 s_2 + \dots + b_m s_m = [b]'[s]$$

In this formula, the biomass regression function is

$$\hat{y} = b_1 x_1 + b_2 x_2 + \dots + b_m x_m = [b]'[x]$$

and the known estimator of the covariance matrix of $[b]$ is $[S_{bb}]$. For example, if the regression function is

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

where d = tree diameter, then, the plot variables s_1 , s_2 and s_3 are the usual plot values of "number of trees per acre", "sum of tree diameters per acre" and "sum of squared tree diameters per acre" respectively.

The estimator of \bar{y}_r (the double sampling with regression estimator when the true values u_h are used), and through \bar{y}_r , the estimator of the parameter of interest μ (the mean biomass per acre) can be defined as

$$w = \bar{u} - c(\bar{v} - \bar{v}^*) = \bar{u} - S_{uv}(\bar{v} - \bar{v}^*)/S_{vv}$$

where

$$\begin{aligned} \bar{u} &= \Sigma \hat{u}_h/n = b_1 \bar{s}_1 + b_2 \bar{s}_2 + \dots + b_m \bar{s}_m \\ &= [b]'[\bar{s}] = \Sigma [b]'[s]/n \end{aligned}$$

and Σ is taken over the n sample plots h of the second phase.

It can be shown that

$$S_{uv} = b_1 S_{s_1 v} + b_2 S_{s_2 v} + \dots + b_m S_{s_m v} = [b]'[S_{sv}]$$

where

$$[S_{sv}]' = [S_{s_1 v} \quad S_{s_2 v} \quad \dots \quad S_{s_m v}]$$

Consequently, w can be successively written as

$$w = [b]'[\bar{s}] - [b]'[S_{sv}](\bar{v} - \bar{v}^*)/S_{vv}$$

$$= [b]'([\bar{s}] - [S_{sv}](\bar{v} - \bar{v}^*)/S_{vv})$$

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

where

$$\begin{aligned} z_1 &= \bar{s}_1 - (S_{s_1 v}/S_{vv})(\bar{v} - \bar{v}^*) = \bar{s}_1 - c_1(\bar{v} - \bar{v}^*) \\ &= \text{double sampling with regression estimator of the mean of } s_1 \end{aligned}$$

$$\begin{aligned} z_2 &= \bar{s}_2 - (S_{s_2 v}/S_{vv})(\bar{v} - \bar{v}^*) = \bar{s}_2 - c_2(\bar{v} - \bar{v}^*) \\ &= \text{double sampling with regression estimator of the mean of } s_2 \end{aligned}$$

$$\begin{aligned} z_3 &= \bar{s}_3 - (S_{s_3 v}/S_{vv})(\bar{v} - \bar{v}^*) = \bar{s}_3 - c_3(\bar{v} - \bar{v}^*) \\ &= \text{double sampling with regression estimator of the mean of } s_3 \end{aligned}$$

etc.

Of course, in these formulae, we have

$$\begin{aligned} c_1 &= S_{s_1 v}/S_{vv} \\ &= \text{estimator of the linear regression coefficient of } s_1 \text{ on } v \end{aligned}$$

$$\begin{aligned} c_2 &= S_{s_2 v}/S_{vv} \\ &= \text{estimator of the linear regression coefficient of } s_2 \text{ on } v \end{aligned}$$

$$\begin{aligned} c_3 &= S_{s_3 v}/S_{vv} \\ &= \text{estimator of the linear regression coefficient of } s_3 \text{ on } v \end{aligned}$$

etc.

Because z_1, z_2, \dots, z_m are double sampling with regression estimators, their variances can be estimated by the formulae

$$\begin{aligned} S_{z_i z_i} &= \left(\frac{n^* - n}{n^*} \right) S_{s_i s_i} | v \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + S_{s_i s_i}/n^* \\ &= \text{estimator of the variance of } z_i, \text{ for } i = 1, 2, \dots, m \end{aligned}$$

where

$$\begin{aligned} S_{s_i s_i} | v &= \Sigma (s_{hi} - s'_{hi})^2 / (n-2) \\ &= (n-1) S_{s_i s_i} (1 - r_i^2) / (n-2) \end{aligned}$$

$$\begin{aligned} s'_{hi} &= (\bar{s}_i - c_i \bar{v}) + c_i v_h \\ &= \text{regression estimate of } s_{hi} \text{ given that } v = v_h \end{aligned}$$

and

$$\begin{aligned} r_i &= S_{s_i v} / \sqrt{S_{s_i s_i} S_{vv}} \\ &= \text{estimator of the linear correlation coefficient of } s_i \text{ and } v \end{aligned}$$

Similarly, the covariance of z_i and z_j , for $i \neq j = 1, 2, \dots, m$ can be estimated by the formula

$$S_{z_i z_j} = \left(\frac{n^* - n}{n^*} \right) S_{s_i s_j} | v \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + S_{s_i s_j}/n^*$$

Where, for Σ meaning summation over clusters h ,

$$\begin{aligned} S_{s_i s_j} | v &= \Sigma (s_{hi} - s'_{hi})(s_{hj} - s'_{hj}) / (n-2) \\ &= (n-1) S_{s_i s_j} (1 - r_i r_j / r_{ij}) / (n-2) \end{aligned}$$

= estimator of the conditional variance of s_i and s_j when v is given,

$$\begin{aligned} s'_{hi} &= (\bar{s}_i - c_i \bar{v}) + c_i v_h \\ &= \text{regression estimate of } s_{hi} \text{ given that } v = v_h \end{aligned}$$

$$s'_{hj} = (\bar{s}_j - c_j \bar{v}) + c_j v_h$$

= regression estimate of s_{hj} given that $v = v_h$

$$r_i = S_{s_i v} / \sqrt{S_{s_i s_i} S_{v v}}$$

= estimator of the linear correlation coefficient of s_i and v

$$r_j = S_{s_j v} / \sqrt{S_{s_j s_j} S_{v v}}$$

= estimator of the linear correlation coefficient of s_j and v

and

$$r_{ij} = S_{s_i s_j} / \sqrt{S_{s_i s_i} S_{s_j s_j}}$$

= estimator of the linear correlation coefficient of s_i and s_j

We have shown above the shortcut, approximate formulae for the calculation of the variance of the double sampling with regression estimator \bar{y}_r . Similarly, for the variance of z_i , $i = 1, 2, \dots, m$ we can write

$$S_{z_i z_i} = S_{s_i s_i} |v| \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1) S_{v v}} \right) + \frac{S_{s_i s_i}}{n^*}$$

when n^* is sufficiently large with respect to n

$$S_{z_i z_i} = (n^* - n) S_{s_i s_i} |v| / n n^* + S_{s_i s_i} / n^*$$

$$= S_{s_i s_i} |v| / n + r_i^2 S_{s_i s_i} / n^*$$

when n is sufficiently large, and

$$S_{z_i z_i} \approx S_{s_i s_i} |v| / n + S_{s_i s_i} / n^* \approx S_{s_i s_i} |v| / n$$

when both n and n^* are sufficiently large and such that $(n^* - n) / n^* \approx 1$ and the effect of the factors $1/n^*$ and $(\bar{v} - \bar{v}^*)^2 / (n-1) S_{v v}$ is negligibly small.

Similarly, one can write the following shortcut approximate formulae for the covariance of z_i and z_j , $i \neq j = 1, 2, \dots, m$

$$S_{z_i z_j} = S_{s_i s_j} |v| \left(\frac{1}{n} + \frac{(\bar{v} - \bar{v}^*)^2}{(n-1) S_{v v}} \right) + \frac{S_{s_i s_j}}{n^*}$$

when n^* is sufficiently large with respect to n

$$S_{z_i z_j} = (n^* - n) S_{s_i s_j} |v| / n n^* + S_{s_i s_j} / n^*$$

$$= S_{s_i s_j} |v| / n + (r_i r_j / r_{ij}) S_{s_i s_j} / n^*$$

when n is sufficiently large, and

$$S_{z_i z_j} \approx S_{s_i s_j} |v| / n + S_{s_i s_j} / n^* \approx S_{s_i s_j} |v| / n$$

when both n and n^* are sufficiently large and such that $(n^* - n) / n^* \approx 1$ and the effect of the factors $(\bar{v} - \bar{v}^*)^2 / (n-1) S_{v v}$ and $1/n^*$ is negligibly small.

As all the elements of the covariance matrix $[S_{zz}]$ of $[z]$ are now defined, the formula for the variance of $w = [b]'[z]$ follows immediately as

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

An Illustrative Example

A forest area has been sampled for biomass in 1960; the trees from 235 randomly selected permanent sample plots were all measured for their diameter values d_1 in inches. Four years later, in 1964 it was desired to update the forest biomass estimates. For this purpose a subsample of 118 of the old 235 plots was selected at random and its trees were measured again for their new diameters d_2 . By defining the 1960, the first measurement plots as the first phase sample and the 1964, the second measurement plots as the second phase sample we can calculate estimators of μ_2 , the mean biomass per acre at the second, 1964 measurement time by the following, double sampling with regression estimator procedures.

Procedure 1: Ignore the error of the biomass regression function and use as auxiliary variable the plot value

v = estimate of the plot biomass per acre at the first, 1960 measurement time.

This is the common, double sampling with regression estimator procedure.

Procedure 2: Take the error of the biomass regression into account and use as auxiliary variable the plot value

v = estimate of the plot biomass per acre at the first, 1960 measurement time.

This is double sampling with (simple linear) regression estimator of the previous section.

Procedure 3: Take the error of the biomass regression into account and use as auxiliary variables the plot values s_1, s_2, \dots, s_m of the first, 1960 measurement time. This is the double sampling with (multiple linear) regression estimator a simple extension of the double sampling with (simple linear) regression estimator of the previous section.

In estimating the tree and plot biomass regression we shall use the biomass regression function calculated by weighted least squares in Cunia (1986b) and already used in Example 1 of Cunia (1986a). For convenience, the statistics of this parabolic regression, that is,

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

$$= b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x]$$

where d = tree diameter (of first or second measurement) and the definitions of x_1, x_2 and x_3 are obvious, are shown below

$$[b]' = [5.1818118 \quad -25.653078 \quad 12.988357]$$

and

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

The use of this regression function implies that the plot variables (that summarize the plot information) are

- s_1 = number of trees per acre
- s_2 = sum of diameters per acre
- s_3 = sum of squared diameters per acre

There are three variables s measured in 1960 on all 235 sample plots and three additional variables s measured in 1964 on the 118 plots of the subsample. Using the first set of three variables s (the 1960 values) one can calculate the variable

$$v = b_1s_1 + b_2s_2 + b_3s_3 = [b]'[s]$$

= 1960 biomass per acre

and using the second set of three variables s (the 1964 values) one can calculate the variable

$$u = b_1s_1 + b_2s_2 + b_3s_3 = [b]'[s]$$

= 1964 biomass per acre

The plot values s_1, s_2, s_3, v and u of the first and second measurement are listed by Cunia (1986d, Tables 1, 2, 3, 7, 8); they are not repeated here. We shall only report the summary values, as they are needed by each procedure.

Procedure 1: The estimator of μ_2 , the mean biomass per acre at the second, 1964 occasion, is the classical double sampling with regression estimator, where the error of the biomass regression function is being ignored, when the error of the estimator \bar{y}_r of μ_2 is being calculated. The variable of interest is u = plot biomass per acre at the second measurement and the auxiliary variable is v = plot biomass per acre at the first measurement. The two variables are highly and positively correlated.

Let us now calculate the necessary statistics. We shall start with the basic sums, sums of squares and sums of cross-products. With Σ meaning summation over the values of the 235 plots measured in 1960 and listed by Cunia (1986d, Tables 7, 8) we find the following sum

$$\Sigma v = 24897913$$

For Σ meaning summation over the 118 permanent sample plot values v and u of the first (1960) and second (1964) measurement respectively listed by Cunia (1986d, Table 8) we find

$$\begin{aligned} \Sigma v &= 12490431, & \Sigma v^2 &= 1969954828000 \\ \Sigma u &= 14084491, & \Sigma u^2 &= 2295905587000 \\ & & \text{and } \Sigma uv &= 2105734088000 \end{aligned}$$

Using the above sums and the notation of the previous section we calculate then

$$\begin{aligned} \bar{v}^* &= \Sigma v/n^* = 24897913/235 = 105948.56 \\ &= \text{estimate of the average biomass per acre at the first, 1960 measurement time; all 235 sample plot data were used.} \end{aligned}$$

$$\begin{aligned} \bar{v} &= \Sigma v/n = 12490431/118 = 105851.11 \\ &= \text{estimate of the average biomass per acre at the first, 1960 measurement time;} \end{aligned}$$

only the data from the 118 permanent plots measured on both occasions were used.

$$\begin{aligned} \bar{u} &= \Sigma u/n = 14084491/118 = 119360.09 \\ &= \text{estimate of the average biomass per acre at the second, 1964 measurement time; only the data from the 118 permanent plots measured on both occasions were used.} \end{aligned}$$

$$\begin{aligned} S_{vv} &= \Sigma(v-\bar{v})^2/(n-1) = 5536998459 \\ &= \text{estimate of the variance of } v_h \text{ based on the 118 permanent plots of the second phase subsample only.} \end{aligned}$$

$$\begin{aligned} S_{uv} &= \Sigma(u-\bar{u})(v-\bar{v})/(n-1) = 5255342828 \\ &= \text{estimate of the covariance of } u_h \text{ and } v_h \end{aligned}$$

$$\begin{aligned} S_{uu} &= \Sigma(u-\bar{u})^2/(n-1) = 5254525850 \\ &= \text{estimate of the variance of } u_h \end{aligned}$$

$$\begin{aligned} c &= S_{uv}/S_{vv} = .94913207 \\ &= \text{estimate of the regression coefficient of the linear regression of } u \text{ on } v \end{aligned}$$

$$\begin{aligned} r^2 &= S_{uv}^2/S_{vv}S_{uu} = .94927965 \\ &= \text{square of the estimate of the linear correlation coefficient of } u \text{ and } v, \text{ and} \end{aligned}$$

$$\begin{aligned} S_{uu|v} &= (n-1) S_{uu}(1-r^2)/(n-2) = 268808928 \\ &= \text{estimate of the (conditional) variance of } u \text{ about the least squares regression line of } u \text{ on } v \end{aligned}$$

We can now calculate the double sampling with (linear) regression estimator of μ_2 , the mean 1964 biomass per acre when (1) the auxiliary variable is the biomass per acre v at the first measurement time in 1960, and (2) the error in the biomass regression function is ignored.

$$\begin{aligned} \bar{y}_r &= \bar{u} - c(\bar{v}-\bar{v}^*) \\ &= 119360.09 - (.94913207)(-97.453529) \\ &= 119360.09 + 92.496270 = 119452.59 \\ &= \text{double sampling with regression estimate of } \mu_2, \text{ the average biomass per acre at the second, 1964 measurement time} \end{aligned}$$

$$\begin{aligned} S_{\bar{y}_r\bar{y}_r} &= \left(\frac{n^*-n}{n^*}\right) S_{uu|v} \left(\frac{1}{n} + \frac{(\bar{v}-\bar{v}^*)^2}{(n-1)S_{vv}}\right) + \frac{S_{uu}}{n^*} \\ &= (.49787234)(268808928)(.0084745909) \\ &\quad + 22359684 \\ &= 23493860 \\ &= \text{estimate of the variance of } \bar{y}_r \end{aligned}$$

$$\sqrt{S_{\bar{y}_r\bar{y}_r}} = \sqrt{23493860} = 4847.0466$$

= estimate of the standard error of \bar{y}_r
and using $t = 2$

$$\begin{aligned} \bar{y}_r \pm t\sqrt{S_{\bar{y}_r\bar{y}_r}} &= (119453 \pm 9694) \text{ pounds} \\ &= 95 \text{ percent confidence limits of } \mu_2 \end{aligned}$$

It may be interesting to see the approximation given by various shortcut formulae.

(1) Although n^* is not large with respect to n , and the value of the correction factor $(n^*-n)/n^* = 117/235$ is approximately equal to .5, we shall use the corresponding variance formula and obtain

$$S_{\bar{y}_r \bar{y}_r} \approx S_{uu|v} \left(\frac{1}{n} + \frac{(\bar{v}-\bar{v}^*)^2}{(n-1)S_{vv}} \right) + \frac{S_{uu}}{n^*} = 24637730$$

The 95 percent confidence limits become

$$(119453 \pm 9927) \text{ pounds}$$

As the reader can verify, the variance is overestimated by

$$\frac{(100)(24637730 - 23493860)}{23493860} = 4.87 \text{ percent}$$

while the 95 percent bound on the error of \bar{y}_r is overestimated by

$$(100)(9927 - 9694)/(9694) = 2.40 \text{ percent}$$

This seems to be an acceptable approximation.

(2) The value of $n = 118$ is sufficiently large and the approximation given by the formula

$$S_{\bar{y}_r \bar{y}_r} \approx \left(\frac{n^*-n}{n^*} \right) S_{uu|v/n} + S_{uu}/n^* \\ = S_{uu|v/n} + r^2 S_{uu}/n^* = 23503635$$

is excellent; the variance of \bar{y}_r is overestimated now by only .04 percent and the 95 percent confidence limits remain practically the same

$$(119453 \pm 9696) \text{ pounds}$$

(3) Using the formula where $(n^*-n)/n^* = 1$, $1/n^* = 0$ and $(\bar{v}-\bar{v}^*)^2/(n-1)S_{vv} = 0$, we have

$$S_{\bar{y}_r \bar{y}_r} = S_{uu|v/n} = 22780417$$

and the 95 percent confidence limits of

$$(119453 \pm 9546) \text{ pounds}$$

practically the same thing as in (2) above.

Procedure 2: To simplify notation, we shall use s_1 , s_2 , and s_3 to denote the variables of the second measurement only; the corresponding variables of the first measurement are only needed to calculate the auxiliary variable v .

By this procedure, the estimate of the average biomass per acre μ_2 , at the second, 1964 measurement time is defined as

$$w = [b]'[z]$$

where

$$z_1 = \bar{s}_1 - c_1(\bar{v}-\bar{v}^*) \\ = \text{double sampling with regression estimator of the population mean "number of trees per acre," say } \mu_{s_1}, \text{ at the second, 1964 measurement time}$$

$$z_2 = \bar{s}_2 - c_2(\bar{v}-\bar{v}^*) \\ = \text{double sampling with regression estimator of the population mean "sum of tree diameters per acre," say } \mu_{s_2}, \text{ at the second, 1964 measurement time}$$

$$z_3 = \bar{s}_3 - c_3(\bar{v}-\bar{v}^*) \\ = \text{double sampling with regression estimator of the population mean "sum of squared tree diameters per acre," say } \mu_{s_3}, \text{ at the second, 1964 measurement time}$$

To calculate z_1 , z_2 , and z_3 , we need the following statistics. For Σ meaning summation over plots $h = 118, 119, \dots, 235$, and using the individual plot values s_1, s_2, s_3 (of the second measurement) and v (of the first measurement) as listed by Cunia (1986d, Tables 2, 8) we calculate the following statistics

$$S_{s_1 s_1} = \Sigma (s_1 - \bar{s}_1)^2 / (n-1) = 24039.555 \\ = \text{estimate of the variance of } s_1$$

$$S_{s_1 s_2} = \Sigma (s_1 - \bar{s}_1)(s_2 - \bar{s}_2) / (n-1) = 123557.78 \\ = \text{estimate of the covariance of } s_1 \text{ and } s_2$$

$$S_{s_1 s_3} = \Sigma (s_1 - \bar{s}_1)(s_3 - \bar{s}_3) / (n-1) = 635935.23 \\ = \text{estimate of the covariance of } s_1 \text{ and } s_3$$

$$S_{s_2 s_2} = \Sigma (s_2 - \bar{s}_2)^2 / (n-1) = 718316.31 \\ = \text{estimate of the variance of } s_2$$

$$S_{s_2 s_3} = \Sigma (s_2 - \bar{s}_2)(s_3 - \bar{s}_3) / (n-1) = 4802637.7 \\ = \text{estimate of the covariance of } s_2 \text{ and } s_3$$

$$S_{s_3 s_3} = \Sigma (s_3 - \bar{s}_3)^2 / (n-1) = 47000205 \\ = \text{estimate of the variance of } s_3$$

$$S_{vv} = \Sigma (v - \bar{v})^2 / (n-1) = 55369985 \\ = \text{estimate of the variance of } v$$

$$S_{s_1 v} = \Sigma (s_1 - \bar{s}_1)(v - \bar{v}) / (n-1) = 4357434.1 \\ = \text{estimate of the covariance of } s_1 \text{ and } v$$

$$S_{s_2 v} = \Sigma (s_2 - \bar{s}_2)(v - \bar{v}) / (n-1) = 41121209 \\ = \text{estimate of the covariance of } s_2 \text{ and } v$$

$$S_{s_3 v} = \Sigma (s_3 - \bar{s}_3)(v - \bar{v}) / (n-1) = 484098884 \\ = \text{estimate of the covariance of } s_3 \text{ and } v$$

$$c_1 = S_{s_1 v} / S_{vv} = .00078696683 \\ = \text{estimate of the linear regression coefficient of } s_1 \text{ on } v$$

$$c_2 = S_{s_2 v} / S_{vv} = .0074266246 \\ = \text{estimate of the linear regression coefficient of } s_2 \text{ on } v$$

$$c_3 = S_{s_3 v} / S_{vv} = .087429839 \\ = \text{estimate of the linear regression coefficient of } s_3 \text{ on } v$$

This yields immediately the values

$$z_1 = 266.01695 \\ - (.00078696683)(105851.11 - 105948.56) \\ = 266.01695 + .076692694 = 266.09364 \\ = \text{double sampling with regression estimate of } \mu_{s_1}, \text{ the mean "number of trees per acre" at the second, 1964 measurement time}$$

$z_2 = 1638.6805$
 $- (.0074266246)(105851.11 - 105948.56)$
 $= 1638.6805 + .72375078 = 1639.4043$
 = double sampling with regression estimate of μ_{s_2} , the mean "sum of tree diameters per acre" at the second, 1964 measurement time

and

$z_3 = 12320.176$
 $- (.087429839)(105851.11 - 105948.56)$
 $= 12320.176 + 8.5203463 = 12328.696$
 = double sampling with regression estimate of μ_{s_3} , the mean "sum of squared diameters per acre" at the second, 1964 measurement time

The variances and covariances of z_1 , z_2 , and z_3 can now be estimated by the formulae

$$S_{z_i z_j} = \left(\frac{n^* - n}{n^*} \right) S_{s_i s_j} | v \left(\frac{1}{n} - \frac{(\bar{v} - \bar{v}^*)^2}{(n-1)S_{vv}} \right) + \frac{S_{s_i s_j}}{n^*}$$

for $i, j = 1, 2, 3$

where, for Σ meaning summation over plots $h = 118, 119, \dots, 235$,

$$S_{s_i s_j} | v = \frac{\sum (s_i - (\bar{s}_i - c_i \bar{v}) - c_i v) (s_j - (\bar{s}_j - c_j \bar{v}) - c_j v)}{(n-2)}$$

$$= (n-1) S_{s_i s_j} (1 - r_i r_j / r_{ij}) / (n-2)$$

$$r_i = S_{s_i v} / \sqrt{S_{s_i s_i} S_{vv}}$$

= estimate of the linear correlation coefficient of s_i and v

and

$$r_{ij} = S_{s_i s_j} / \sqrt{S_{s_i s_i} S_{s_j s_j}}$$

= estimate of the linear correlation coefficient of s_i and s_j

Obviously, for any $i = 1, 2, 3$, we have $r_{ii} = 1$

Doing the necessary algebraic calculations we find first

$$S_{s_1 s_1} | v = 20788.075 \quad S_{s_1 s_2} | v = 91982.936$$

$$S_{s_1 s_3} | v = 257163.45 \quad S_{s_2 s_2} | v = 416484.22$$

$$S_{s_2 s_3} | v = 1217825.8 \quad \text{and} \quad S_{s_3 s_3} | v = 4715823.6$$

Then, we calculate the covariances of z_i and z_j and arrange them in the following covariance matrix

$$[S_{zz}] = \begin{bmatrix} 102.29583 & 525.77714 & 2706.1055 \\ 525.77714 & 3056.6621 & 20436.747 \\ 2706.1055 & 20436.747 & 200000.84 \end{bmatrix}$$

Consequently,

$w = [b]'[z] = 119452.59$
 = double sampling with regression estimate of μ_2 , the average biomass per acre at the second, 1964 measurement time

$S_{ww} = [b]'[S_{zz}][b] + [z]'[S_{bb}][z]$
 $= 22359683 + 11407373 = 33767056$
 = estimate of the variance of w

$\sqrt{S_{ww}} = \sqrt{33767056} = 5810.9427$
 = estimate of the standard error of w

and using a t-value of 2,

$$w \pm t\sqrt{S_{ww}} = (119453 \pm 11622) \text{ pounds}$$

= 95 percent confidence limits of μ_2

As the reader can verify, ignoring the effect of the error of biomass tables, results in an underestimation of the variance of w by

$$(100)(33767056 - 23493860) / (33767056) = 30.42 \text{ percent}$$

The standard error of w is underestimated by

$$(100)(5810.9427 - 4847.0466) / (5810.9427) = 16.59 \text{ percent}$$

While the estimate of μ_2 is exactly the same by the two procedures, the estimate of the variance of w by procedure 1 is slightly higher than the value of the first component of the variance as estimated by the procedure 2. The difference is small, only

$$(100)(23493860 - 22359683) / (23493860) = 4.83 \text{ percent}$$

Procedure 3: It is felt that the regression estimators of the mean of s_1 , s_2 and s_3 (of the second measurement time) of Procedure 2 above may be improved if the auxiliary variable v is replaced by the first measurement values of the variables s_1 , s_2 and s_3 . Recall that v is a linear combination of the first measurement values of s_1 , s_2 and s_3 . To simplify notation we shall use

- (1) s_1 , s_2 and s_3 to denote the variables of the second measurement and
- (2) $v_1 = s_1$ of the first measurement
 $v_2 = s_2$ of the first measurement, and
 $v_3 = s_3$ of the first measurement.

The values v_1 , v_2 and v_3 are to be found in Cunia (1986d, Tables 1, 2) and the values s_1 , s_2 and s_3 in Cunia (1986d, Table 2). Using the first measurement values $v_1 = s_1$, $v_2 = s_2$ and $v_3 = s_3$ of the 235 plots that were measured in 1964 we find

$$[\bar{v}^*] = \begin{bmatrix} \bar{v}_1^* \\ \bar{v}_2^* \\ \bar{v}_3^* \end{bmatrix} = \begin{bmatrix} 225.61702 \\ 1389.0374 \\ 10810.647 \end{bmatrix}$$

= vector of the average values of the variables v_1 , v_2 and v_3 as calculated from the first measurement data of the entire sample of 235 plots

Using now the first and the second measurement values of the subsample of 118 permanent plots measured in 1960 and 1964 (variables v_1 , v_2 and v_3 for the first, and variables s_1 , s_2 and s_3 for the second measurement) we find the following statistics

$$[\bar{v}] = \begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \bar{v}_3 \end{bmatrix} = \begin{bmatrix} 236.35593 \\ 1439.3161 \\ 10898.164 \end{bmatrix}$$

= vector of the sample averages of v_1 , v_2 and v_3 of the 118 permanent plots

$$[S_{VV}] = \begin{bmatrix} 21272.932 & 110452.37 & 602351.30 \\ 110452.37 & 659011.32 & 4724541.6 \\ 602351.30 & 4724541.6 & 48604076 \end{bmatrix}$$

= estimate of the covariance matrix of v_1 , v_2 and v_3 as calculated from the data of the 118 permanent plots

$$[\bar{s}] = \begin{bmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \bar{s}_3 \end{bmatrix} = \begin{bmatrix} 266.01695 \\ 1638.6805 \\ 12320.176 \end{bmatrix}$$

= vector of the sample averages of s_1 , s_2 and s_3 of the 118 permanent plots

$$[S_{SS}] = \begin{bmatrix} 24039.555 & 123557.78 & 635935.23 \\ 123557.78 & 718316.31 & 4802637.7 \\ 635935.23 & 4802637.7 & 47000205 \end{bmatrix}$$

= estimate of the covariance matrix of s_1 , s_2 and s_3 as calculated from the data of the 118 permanent plots

$$[S_{SV}] = \begin{bmatrix} 21987.926 & 109292.82 & 542577.80 \\ 118295.97 & 670898.34 & 4443890.0 \\ 659678.73 & 4832348.2 & 46552860 \end{bmatrix}$$

= estimate of the covariance matrix of s_1 , s_2 and s_3 with v_1 , v_2 and v_3 as calculated from the data of the 118 permanent plots

Note that the first row of $[S_{SV}]$ contains the covariances of s_1 with v_1 , v_2 , v_3 , the second column of $[S_{SV}]$ contains the covariances of v_2 with s_1 , s_2 and s_3 , etc.

Using the above values we calculate

$$[C] = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = [S_{VV}]^{-1} [S_{SV}]'$$

$$= \begin{bmatrix} 1.7649801 & 3.5411706 & 9.0746818 \\ -1.7546979 & .27603189 & -.81992294 \\ .0063462595 & .020713020 & .92503505 \end{bmatrix}$$

= estimate of the matrix of regression coefficients of s_1 (c_{11} , c_{21} , c_{31}), s_2 (c_{12} , c_{22} , c_{32}) and s_3 (c_{13} , c_{23} , c_{33}) as the dependent variables and v_1 , v_2 and v_3 as the independent variables.

Note that the regression function of s_1 on v_1 , v_2 and v_3 can be written as

$$s_1' = \bar{s}_1 + c_{11}(v_1 - \bar{v}_1) + c_{21}(v_2 - \bar{v}_2) + c_{31}(v_3 - \bar{v}_3)$$

$$= 32.247354 + 1.7649801 v_1 - .17546979 v_2 + .0063462595 v_3$$

with similar expressions for the regression functions of s_2 and s_3 .

We can now calculate the vector of statistics $[z]$ by the formula

$$[z] = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = [\bar{s}] - [C]'[\bar{v} - \bar{v}^*] = \begin{bmatrix} 255.32996 \\ 1584.9609 \\ 12182.992 \end{bmatrix}$$

To estimate the covariance matrix of $[z]$ we need to calculate first the matrix $[R]$ of the residuals of s_1 , s_2 and s_3 from their own regressions. The first column of $[R]$ contains the 118 residuals ($s_1 - s_1'$) defined as

$$s_1 - s_1' = (s_1 - \bar{s}_1) - c_{11}(v_1 - \bar{v}_1) - c_{21}(v_2 - \bar{v}_2) - c_{31}(v_3 - \bar{v}_3)$$

Defined by similar formulae, the two sets of 118 residuals ($s_2 - s_2'$) and ($s_3 - s_3'$) are contained in the second and third column respectively of $[R]$. Dividing the product $[R]'[R]$ by the number $(n-4) = 114$ degrees of freedom yields the covariance matrix

$$[S_{SS|V}] = [R]'[R]/114$$

$$= \begin{bmatrix} 990.96189 & 4400.9063 & 24744.489 \\ 4400.9063 & 22757.914 & 172903.71 \\ 24744.489 & 172903.71 & 1963296.7 \end{bmatrix}$$

= estimate of the conditional covariance matrix of the variables s_1 , s_2 and s_3 for given v_1 , v_2 and v_3 , calculated from the data of the 118 plots of the second phase subsample.

This yields the statistic

$$[S_{ZZ}] = \left(\frac{n^* - n}{n^*} \right) \left([\bar{v} - \bar{v}^*]' [S_{VV}]^{-1} [\bar{v} - \bar{v}^*] \right) [S_{SS|V}] + \frac{1}{n^*} [S_{SS}]$$

$$= \begin{bmatrix} 108.87373 & 554.98988 & 2870.3549 \\ 554.98988 & 3207.7263 & 21584.446 \\ 2870.3549 & 21584.446 & 213032.73 \end{bmatrix}$$

= estimate of the covariance

The point and interval estimates of the average biomass per acre μ_2 at the second, 1964 measurement time follows immediately as

$$w = [b]'[z] = 118900.99 = \text{double sampling with (multiple linear) regression estimate of } \mu_2$$

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

$$= 23907253 + 11798043 = 35705297$$

= estimate of the variance of w

$$\sqrt{S_{ww}} = \sqrt{35705297} = 5975.3909$$

= estimate of the standard error of w

and using a t-value of 2,

$$w \pm 2\sqrt{S_{ww}} = (118901 \pm 11951) \text{ pounds}$$

= 95 percent confidence limits of μ_2

The reader may be surprised by the fact that the error of the estimate by Procedure 2 (using simple linear regression) is smaller than the error by Procedure 3 (that uses multiple linear regression). The standard error of w is 5811 by Procedure 2 and 5975 by Procedure 3, an increase of some 2.8 percent. This may be due to sampling error. A more credible reason is, however, that the multiple linear regression of s_1 on v_1 , v_2

and v_3 may not necessarily be better than the simple linear regression of s_i on their linear combination $v = b_1v_1 + b_2v_2 + b_3v_3$. The error of the multiple linear regression has four sources (the four regression coefficients) compared to the error of the simple linear regression that has only two (the two regression coefficients). Consequently, whenever the multiple linear regression is not significantly better, it may well be possible for the corresponding double sampling estimator to be less precise than the double sampling with simple linear regression estimator.

Acknowledgements

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

Literature Cited

- Cochran, W. G. Sampling Techniques, 3rd Ed. John Wiley and Sons, New York, NY; 1977.
- Cunia, T. Some theory on the reliability of volume estimates in a forest inventory sample. Forest Science, 11:115-128; 1965.
- Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.
- Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986c.
- Cunia, T. On the error of biomass estimates in forest inventories: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (86-001), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

ON THE ERROR OF FOREST INVENTORY ESTIMATES:

CONTINUOUS FOREST INVENTORY WITHOUT SPR⁽¹⁾

Tiberius Cunia

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

to the parameter of interest μ , say $\mu = [\beta]'[\mu_z]$. The variance of w can be estimated by the approximate formula

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

where $[S_{zz}]$ and $[S_{bb}]$ are the estimates of the covariance matrices of $[z]$ and $[b]$ respectively. Note that the first part of S_{ww} may be viewed as the error component due to sample plots and the second part may be viewed as the error component due to biomass regressions.

The definition of $[z]$ depends on the (i) specific parameter μ one wishes to estimate, (ii) specific sampling design by which the sample plots are selected (including the type of sample units used) and (iii) specific independent variables x_1, x_2, \dots, x_m used in the biomass regression function. In the Cunia (1986a) paper it is assumed that (i) the parameter to estimate is the average biomass per acre μ , (ii) the sample plots (or points) are selected by simple random sampling and (iii) the definition of the statistics z is based on the plot (or point) variables s_1, s_2, \dots, s_m defined as the averages of the regression variables x_1, x_2, \dots, x_m expressed on a "per acre" basis.

For example, let us assume that $x_1 = 1$, $x_2 = d = \text{tree diameter}$ and $x_3 = d^2$. Then, for Σ meaning summation over the trees of a given plot, we have the plot variables

$$\begin{aligned} s_1 &= (\Sigma x_1) / (\text{plot area}) \\ &= \text{number of trees per acre} \\ s_2 &= (\Sigma x_2) / (\text{plot area}) \\ &= \text{sum of tree diameters per acre, and} \\ s_3 &= (\Sigma x_3) / (\text{plot area}) \\ &= \text{sum of squared diameters per acre} \end{aligned}$$

If instead of a plot we have a relascope point sample, the variables s_1, s_2 and s_3 are similarly defined; only the way they are calculated is different. For more details on this, the reader is referred to Cunia (1986a). If, in addition, the plots are selected by simple random sampling and the parameter of interest is the average biomass per acre μ , the variables z_1, z_2 and z_3 are defined as the sample averages \bar{s}_1, \bar{s}_2 and \bar{s}_3 respectively. For this case, the covariance matrix $[S_{zz}]$ of $[z]$ is easy to calculate; its element ij is simply the sample covariance of \bar{s}_i and \bar{s}_j , that is

$$\begin{aligned} S_{z_i z_j} &= S_{s_i s_j} / n \\ &= \Sigma (s_i - \bar{s}_i)(s_j - \bar{s}_j) / n(n-1) \end{aligned}$$

where n is the number of sample plots and Σ is taken over the n sample plots.

It is the objective of the present paper to extend the application of this approach to biomass estimates calculated from Continuous Forest Inventory (CFI) data of two successive measurements. The Sampling with Partial Replacement (SPR) is not being used; the case of CFI with SPR is considered in a companion paper by Cunia (1986b). We shall consider estimates of the current average biomass

The error of the biomass estimates in Continuous Forest Inventory Systems has two main error components; one due to sample plots and one due to biomass tables or regressions. The common procedures by which the estimates are calculated take into account only the first component; the second component is simply ignored. An approach is shown that introduces the error of the biomass regressions into the total error of the estimates of average biomass per tree, average biomass per acre and growth components such as average net change from one to the next occasion, average mortality or average ingrowth biomass per acre.

Introduction

The sampling design of a forest inventory system consists generally of a random selection of sample plots (or Bitterlich relascope points) where the trees are measured for diameter and where the biomass of these sample trees is estimated by biomass tables or regression functions. When inferences about the error of the forest biomass estimates are made, however, only the error from the sample plots (or points) is taken into account; the error of the biomass tables or regressions is simply ignored.

Cunia (1986a) has proposed an approach to combine the error of the biomass regressions with that of the sample plots (or points). This approach requires that the estimators be of the form

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

where $[b]$ is the estimate of the vector of coefficients of the biomass regression

$$\hat{y} = [b]'[x] = b_1 x_1 + b_2 x_2 + \dots + b_m x_m$$

y is the tree biomass and $[z]$ is the vector of statistics calculated from the sample plot data. We implicitly assume here that the statistics z are defined so that the product of the expected value of the vector $[z]$, say $[\mu_z]$ with the expected value of the vector $[b]$, say $[\beta]$ is equal

(1) Paper based on a set of lecture notes "On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots" Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

per unit area (on the first and second measurement occasion), estimate of the average biomass per tree and estimates of growth components per unit area such as mortality, ingrowth, growth on survivor trees and net changes in biomass between the two measurements. As the difference between working with sample plots or Bitterlich sample points rests with the way the variables s_1, s_2, \dots, s_m are calculated, we shall implicitly assume that whatever we say for sample plots apply equally well to sample points.

The CFI sample units are generally selected by systematic sampling but we shall assume that the bias due to the application of the simple random sampling formulae is negligibly small. It can be shown that this bias affects only the variance component associated with the sample plots. We shall also assume that the biomass tables are constructed from linear regression functions $\hat{y} = [b]'[x]$ for which the estimates of the covariance matrices $[S_{bb}]$ are given. For ways to calculate $[b]$ and $[S_{bb}]$ the reader is referred to Cunia (1986c,d) among others. The non-statistical error of using, in a given forest area, biomass tables constructed for another area is not being considered here.

We shall specifically define the vectors $[z]$ corresponding to each CFI estimator and show how to estimate their covariance matrices. Once $[z]$ and $[S_{zz}]$ are defined, the estimators w and S_{ww} can be calculated by the formulae given above. Finally, we shall illustrate the application of these formulae to a simple numerical example.

In what follows we shall use the notation

$$\hat{y}_{hk} = b_1 x_{hk1} + b_2 x_{hk2} + \dots + b_m x_{hkm}$$

= $[b]'[x_{hk}]$ = regression estimate of the true but unknown biomass y_{hk} of the k -th tree of the h -th plot

and, for the case of sample plots of fixed area a ,

$$\hat{v}_h = (\Sigma \hat{y}_{hk})/a = \Sigma (\hat{y}_{hk}/a)$$

= regression estimate of the true but unknown biomass per acre v_h of plot h ,

where Σ is taken over the trees of plot h .

When the sample unit is a Bitterlich sample point, the definition of \hat{v}_h becomes

$$\hat{v}_h = \Sigma (\hat{y}_{hk}/a_{hk})$$

$$= (b_1 s_{h1} + b_2 s_{h2} + \dots + b_m s_{hm})$$

where a_{hk} is the plot area corresponding to the hk -th tree, and

$$s_{hi} = \Sigma (x_{hki}/a_{hk}) \text{ for } i = 1, 2, \dots, m$$

is the "sum of x_i values per unit area" at the sample point h .

For convenience, we may prefer using sometimes the more general notation

$$\hat{v}_h = \Sigma (\hat{y}_{hk}/a_{hk})$$

for both plots of fixed area (where $a_{hk} = a$ for all trees) and Bitterlich sample points. For more details on these variables the reader is sent to Cunia (1986a).

Because we shall be dealing with variables s measured on two occasions, it may be convenient to use a superscript 1 for the first and 2 for the second measurement time. For example, s_2^1 will denote variable s_2 measured on the first occasion, while s_1^2 will denote variable s_1 measured on the second occasion, and not the square of s_1 . We realize that, sometimes, this may lead to confusion. However, we shall remind the reader, whenever we feel it necessary, whether 2 is a superscript or exponent. Of course, when the superscript is not necessary, we shall simplify the notation by dropping it.

Biomass Estimators

We shall assume that (i) the CFI sample consists of n plots selected by simple random sampling, (ii) the biomass table is constructed from a linear regression function for which $[b]$ and $[S_{bb}]$ are given, and (iii) the data of the sample plots, the variables s , are statistically independent of the biomass tables.

Estimator of the Average Biomass per Unit Area μ_1 at the First Measurement Time

As the discussion is limited to the first measurement data, superscript 1 will be dropped from the notation of s_h but used only in the definition of $[z]$ and its covariance matrix $[S_{zz}]$. For notational convenience we shall drop the subscript h of the variables s_{hi} and let Σ stand for summation over the values of the plot $h = 1, 2, \dots, n$.

The common procedure is to define the estimator of μ_1 as

$$w_1 = \Sigma \hat{v}_h / n = (b_1 \Sigma s_1 + b_2 \Sigma s_2 + \dots + b_m \Sigma s_m) / n$$

$$= b_1 \bar{s}_1 + b_2 \bar{s}_2 + \dots + b_m \bar{s}_m = [b]'[\bar{s}]$$

If we define now

$$z_i^1 = \bar{s}_i = \Sigma s_i / n \text{ for } i = 1, 2, \dots, m$$

we can define $[z^1] = [\bar{s}]$ and, thus,

$$w_1 = [b]'[z^1] = \text{estimator of } \mu_1$$

Because $[b]$ and $[z^1]$ are statistically independent the variance of w_1 can be estimated by the approximate formula

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

where

$$[S_{zz}^{11}] = \text{estimator of the covariance matrix of } [z^1].$$

It remains now to calculate $[S_{zz}^{11}]$. For this we use the fact that, for $i, j = 1, 2, \dots, m$,

$$(1) \quad S_{s_i s_j} = \Sigma (s_i - \bar{s}_i)(s_j - \bar{s}_j) / (n-1)$$

= estimator of the covariance of s_i and s_j

and

$$(2) \quad S_{\bar{s}_i \bar{s}_j} = S_{s_i s_j} / n = S_{z_i z_j}$$

If

$[S_{ss}^{11}]$ = estimator of the covariance matrix of s_1, s_2, \dots, s_m as calculated from the first measurement values,

we can verify immediately that

$$[S_{zz}^{11}] = [S_{ss}^{11}] / n$$

Estimator of the Average Biomass per Unit Area μ_2 at the Second Measurement Time

The procedure is the same as in the subsection above with the second, rather than the first measurement values s_1, s_2, \dots, s_m being used

If

$[z^2]$ = $[\bar{s}]$ = vector of the average values of s_1, s_2, \dots, s_m as measured on the second occasion,

and

$[S_{zz}^{22}] = [S_{ss}^{22}] / n$ = estimator of the covariance matrix of $[z^2]$

where

$[S_{ss}^{22}]$ = estimator of the covariance matrix of s_1, s_2, \dots, s_m as calculated from the second measurement data,

we can verify immediately that

$$w_2 = [b]' [z^2] = \text{estimator of } \mu_2$$

and

$$S_{w_2 w_2} = [b]' [S_{zz}^{22}] [b] + [z^2]' [S_{bb}] [z^2]$$

= estimator of the variance of w_2

Note that in the formulae above, 2 is a superscript not an exponent.

Estimator of the Net Change $\mu_g = (\mu_2 - \mu_1)$ in Biomass per Unit Area between the Two Measurements

The estimator of the net change can be simply defined as

$$w_g = w_2 - w_1 = [b]' [z^2] - [b]' [z^1] = [b]' [z^g]$$

where

$[z^g] = [z^2] - [z^1]$ = difference between the sample averages of $[s]$ of the first and second measurement

Because the i -th element of $[z^g]$, $i = 1, 2, \dots, m$, can be written as

$$\bar{s}_i^2 - \bar{s}_i^1 = \Sigma (s_{hi}^2 - s_{hi}^1) / n = \Sigma s_{hi}^g / n = \bar{s}_i^g$$

it suffices to work with the individual cluster values s_{hi}^g , define their averages \bar{s}_i^g and covariances $S_{s_i s_i}^{gg}$, arranged as the vector of averages $[S^g]$ and covariance matrix $[S_{ss}^{gg}]$.

Then

$$[z^g] = [S^g] \text{ and } [S_{zz}^{gg}] = [S_{ss}^{gg}] / n$$

and

$$w_g = [b]' [z^g] = \text{estimator of } \mu_g, \text{ and}$$

$$S_{w_g w_g} = [b]' [S_{zz}^{gg}] [b] + [z^g]' [S_{bb}] [z^g]$$

Estimators of Growth Components: μ_s = Average Biomass Growth per Unit Area on Survivor Trees, μ_m = Average Mortality Biomass per Unit Area and μ_i = Average Ingrowth Biomass per Unit Area

When μ_g was calculated, all trees measured on either occasion were used. If instead of all trees one uses only the survivor trees, that is, the trees that were measured on the first and were alive and measured also on the second occasion, one can define the estimator w_s of μ_s . More specifically, we define the new variables of plot h as

$$s_{hi}^s = \Sigma ((x_{hki}^2 - x_{hki}^1) / a_{hk}) \text{ for } i = 1, 2, \dots, m$$

where Σ is taken only over the trees k of plot h that were alive and measured on both occasions. Note that x^2 denotes the value of x as measured on the second occasion and not the square of x . For the case of the relascope sample points, the "per unit area" converting factor a_{hk} is that of the first measurement. We further define

$[z^s] = [\bar{s}^s]$ = vector of sample averages \bar{s}_i^s of s_{hi}^s

$[S_{ss}^{ss}]$ = sample covariance matrix of the new plot variables s_{hi}^s

and

$[S_{zz}^{ss}] = [S_{ss}^{ss}] / n$ = estimator of the covariance matrix of $[z^s]$.

This yields immediately

$$w_s = [b]'[z^s] = \text{estimator of } \mu_s$$

and

$$S_{w_s w_s} = [b]'[S_{zz}^{ss}][b] + [z^s]'[S_{bb}^s][z^s]$$

If in w_1 , the estimator of the average biomass per unit area of the first measurement we use only the mortality trees (the trees that were measured on the first occasion but not measured on the second because they were dead, not harvested) we obtain the estimator w_m of μ_m . More specifically, the new plot variables are

$$s_{hi}^m = \Sigma(x_{hki}^1/a_{hk}) \text{ for } i = 1, 2, \dots, m$$

where Σ is taken only over the mortality trees k of plot h. Using these new plot variables we further define

$$[z^m] = [\bar{s}^m] = \text{vector of sample averages } \bar{s}_i^m$$

$$[S_{ss}^{mm}] = \text{sample covariance matrix of } s_{hi}^m$$

$$[S_{zz}^{mm}] = [S_{ss}^{mm}]/n = \text{estimator of the covariance matrix of } [z^m]$$

$$w_m = [b]'[z^m] = \text{estimator of } \mu_m, \text{ and}$$

$$S_{w_m w_m} = [b]'[S_{zz}^{mm}][b] + [z^m]'[S_{bb}^m][z^m]$$

= estimator of the variance of w_m

Similarly, if in w_2 , the estimator of the average biomass per unit area at the second occasion, we use only the ingrowth trees (the trees that were too small to be measured on the first but became of merchantable size on the second occasion) we obtain an estimator w_i of μ_i . More specifically, we define the new plot variables

$s_{hi}^i = \Sigma(x_{hki}^2/a_{hk})$ for subscript $i = 1, 2, \dots, m$ (superscript i means ingrowth) where Σ is now taken only over the ingrowth trees k of plot h. For the case of relascope sample points, the per "unit area" factor a_{hk} is that of the second measurement time. Using the new plot variables we define the usual statistics

$$[z^i] = [\bar{s}^i] = \text{vector of sample averages } \bar{s}_i^i$$

$$[S_{ss}^{ii}] = \text{sample covariance matrix of } s_{hi}^i$$

$$[S_{zz}^{ii}] = [S_{ss}^{ii}]/n = \text{estimator of the covariance matrix of } [z^i]$$

$$w_i = [b]'[z^i] = \text{estimator of } \mu_i, \text{ and}$$

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

= estimator of the variance of w_i

We hope that use of superscripts s , m and i to denote survivor, mortality and ingrowth trees will not be confused with the notation s for plot variables, m for the number of variables x or s , and i for subscript of variables x_i or s_i .

Note that the estimators w_s , w_m and w_i are,

most probably biased. This is because the biomass regression function is usually defined for the entire set of all population trees. When applied to a class of trees, it is not uncommon for the biomass of the trees in this class to fall, on the average, below or above the overall regression function. For example, it is reasonable to expect the survivor trees to be the tallest and most vigorous trees of highest growth (the dominant or codominant trees) and the mortality trees to be the shortest, least vigorous trees (intermediate or suppressed). Thus, it is reasonable to assume that w_s is an underestimate and w_m an overestimate. But there is also a source of underestimation of w_m ; the growth on mortality trees between their measurement on the first occasion and the time of their death is being ignored.

It can be shown that, in the absence of harvesting, we must have (for plot, not point sampling)

$$w_2 = w_1 + w_s + w_i - w_m \text{ and } w_g = w_s + w_i - w_m$$

These relationships can be used as a check on the calculations performed.

Estimator of the Average Biomass per Tree μ_t at a Given Occasion

Because the estimation of the average biomass per tree follows the same procedure for either measurement, we shall consider the general case with no need of superscript 1 or 2.

While the plots (clusters of trees) are selected by simple random sampling, the selection of the trees themselves is made by cluster random sampling. To estimate μ_t we must therefore use the cluster sampling formulae as found in standard texts on sampling techniques as, for example, that of Cochran (1977). Using our notation, these formulae can be written as follows

$$w_t = \Sigma \hat{v}_h / \Sigma t_h = \text{estimator of } \mu_t, \text{ and}$$

$$S_{w_t w_t} = n(S_{vv} - 2w_t S_{vt} + w_t^2 S_{tt}) / (\Sigma t_h)^2$$

$$= n(\Sigma v_h^2 - 2w_t \Sigma v_h t_h + w_t^2 \Sigma t_h^2) / (n-1)(\Sigma t_h)^2$$

= estimator of the variance of w_t

where Σ is taken over all plots $h = 1, 2, \dots, n$ and t_h is used to denote the number of trees per unit area of plot h . Do not confuse subscript t of w_t that stands for "biomass per tree" (and may also be used as superscript) and variable t that denotes the number of trees per unit area (which may also be used as a subscript).

In this formula we have ignored the effect of the finite population correction factor; the population size is ordinarily much larger than the sample size. We have also ignored the effect of the error of the biomass tables. To take this error into account we must express first the plot biomass per unit area as

$$\hat{v}_h = b_1 s_{h1} + b_2 s_{h2} + \dots + b_m s_{hm} = [b]'[s_h]$$

Algebraic manipulation of the formula of w_t yields the formula

$$w_t = b_1 z_1 + b_2 z_2 + \dots + b_m z_m = [b]' [z^t]$$

where

$z_i = \Sigma s_{hi} / \Sigma t_h =$ cluster sampling estimator of the expected value of the tree variable x_i .

The variance of z_i is estimated by the formula

$$S_{z_i z_i} = n(S_{s_i s_i} - 2z_i S_{s_i t} + z_i^2 S_{tt}) / (\Sigma t_h)^2$$

$$= n(\Sigma s_{hi}^2 - 2z_i \Sigma s_{hi} t_h + z_i^2 \Sigma t_h^2) / (n-1)(\Sigma t_h)^2$$

and the covariance of z_i and z_j by the formula

$$S_{z_i z_j} = n(S_{s_i s_j} - z_i S_{s_j t} - z_j S_{s_i t} + z_i z_j S_{tt}) / (\Sigma t_h)^2$$

$$= n(\Sigma s_{hi} s_{hj} - z_i \Sigma s_{hj} t_h - z_j \Sigma s_{hi} t_h + z_i z_j \Sigma t_h^2) / (n-1)(\Sigma t_h)^2$$

Using the formulae above, we calculate

$$[S_{zz}^{tt}] = \text{estimator of the covariance matrix of } [z^t]$$

This yields immediately

$$w_t = [b]' [z^t] = \text{estimator of the average biomass per tree}$$

and

$$S_{w_t w_t} = [b]' [S_{zz}^{tt}] [b] + [z^t] [S_{bb}] [z^t]$$

= estimator of the variance of w_t

It may be interesting to see what happens when $z_i \equiv 1$. This may be the case when the sample unit is a plot of fixed area "a"

and

$$s_{h1} = \Sigma (1/a) = \text{number of trees per unit area} = t_h$$

where Σ is taken over all trees of a given plot h.

Then,

$$z_1 = \Sigma s_{h1} / \Sigma t_h \equiv 1$$

and it can be easily shown that, for all $i = 1, 2, \dots, m$

$$S_{z_1 z_i} = 0$$

Illustrative Numerical Example

To see how to apply the formulae of the previous section, we shall work with data from 118 permanent, one-tenth acre sample plots selected at random from the Maniwaki region of Québec-

Canada. These plots were measured on two occasions in 1960 and 1964. To estimate the biomass of sample trees and plots, we shall use a tree biomass regression function of the parabolic form

$$y = b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]' [x]$$

where $x_1 = 1$, $x_2 =$ (tree diameter) and $x_3 =$ (squared tree diameter). This regression function was calculated from a random sample of 353 trees by Cunia (1986c) and the following statistics are given.

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

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

The biomass is given in pounds of above ground green weight and the tree diameter is measured in inches.

The plot variables corresponding to this regression function are

$s_1 =$ number of trees/acre

$s_2 =$ sum of tree diameters/acre, and

$s_3 =$ sum of squared tree diameters/acre

Sometimes, a superscript 1 for the first and 2 for the second measurement may be added. It should not be confused with exponents 1 or 2.

Estimation of μ_1 , μ_2 and $\mu_g = (\mu_2 - \mu_1)$

The values of the individual plot variables s_1 , s_2 and s_3 as measured in 1960 and 1964 on the $n = 118$ permanent plots of our sample are given in Cunia (1986e), and the values of their means, variances and covariances are shown below.

$$[s^1] = \begin{bmatrix} 236.35593 \\ 1439.3161 \\ 10898.164 \end{bmatrix} = \text{sample mean of } [s] \text{ of the first measurement}$$

$$[s^2] = \begin{bmatrix} 266.01695 \\ 1638.6805 \\ 12320.176 \end{bmatrix} = \text{sample mean of } [s] \text{ of the second measurement}$$

$$[S_{ss}^{11}] = \begin{bmatrix} 21272.932 & 110452.37 & 602351.30 \\ 110452.37 & 659011.32 & 4724541.6 \\ 602351.30 & 4724541.6 & 48604076 \end{bmatrix}$$

= sample covariance matrix of $[s]$ of the first measurement

$$[S_{ss}^{22}] = \begin{bmatrix} 24039.555 & 123557.78 & 635935.23 \\ 123557.78 & 718316.31 & 4802637.7 \\ 635935.23 & 4802637.7 & 47000205 \end{bmatrix}$$

= sample covariance matrix of $[s]$ of the second measurement

$$[S_{ss}^{12}] = \begin{bmatrix} 21987.926 & 118295.97 & 659678.73 \\ 109292.82 & 670898.34 & 4832348.2 \\ 542577.80 & 4443890.0 & 46552860 \end{bmatrix}$$

= sample covariance matrix of [s] of the first with [s] of the second measurement

Note that $[S_{ss}^{12}]$ is not symmetrical and that the covariance terms associated with s_1 of the first measurement are found in row 1, while the covariance terms associated with s_1 of the second measurement are found in column 1.

These statistics are sufficient to estimate μ_1 , μ_2 and net change μ_g and their variances. Using the formulae of the previous section, the reader can verify that

$$[z^1] = [\bar{s}^1] = \begin{bmatrix} 236.35593 \\ 1439.3161 \\ 10898.164 \end{bmatrix},$$

$$[z^2] = [\bar{s}^2] = \begin{bmatrix} 266.01695 \\ 1638.6805 \\ 12320.176 \end{bmatrix}$$

$$[z^g] = [z^2] - [z^1] = \begin{bmatrix} 29.661017 \\ 199.36441 \\ 1422.0114 \end{bmatrix}$$

$$[S_{zz}^{11}] = [S_{ss}^{11}]/n = \begin{bmatrix} 180.27909 & 936.03701 & 5104.6720 \\ 936.03701 & 5584.8417 & 40038.488 \\ 5104.6720 & 40038.488 & 411898.95 \end{bmatrix}$$

$$[S_{zz}^{22}] = [S_{ss}^{22}]/n = \begin{bmatrix} 203.72504 & 1047.0999 & 5389.2817 \\ 1047.0999 & 6087.4263 & 40700.319 \\ 5389.2817 & 40700.319 & 398306.82 \end{bmatrix}$$

and

$$[S_{zz}^{gg}] = \left([S_{ss}^{11}] - [S_{ss}^{12}] - [S_{ss}^{12}]' + [S_{ss}^{22}] \right) / n$$

$$= \begin{bmatrix} 11.327426 & 54.418338 & 305.33909 \\ 54.418338 & 301.10981 & 2126.6194 \\ 305.33909 & 2126.6194 & 21174.239 \end{bmatrix}$$

This yields immediately the estimates

$$w_1 = [b]'[z^1] = 105851.11 \text{ pounds}$$

= estimate of the mean biomass per acre at the first, 1960 measurement

$$w_2 = [b]'[z^2] = 119360.09 \text{ pounds}$$

= estimate of the mean biomass per acre at the second, 1964 measurement

$$w_g = [b]'[z^g] = 13508.98 \text{ pounds} = w_2 - w_1$$

= estimate of the mean change in biomass per acre between the two measurements

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

$$= 46923716 + 10710584 = 57634300$$

= estimate of the variance of w_1

$$S_{w_2 w_2} = [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2]$$

$$= 44529880 + 11393835 = 55923716$$

= estimate of the variance of w_2

and

$$S_{w_g w_g} = [b]'[S_{zz}^{gg}][b] + [z^g]'[S_{bb}][z^g]$$

$$= 2379986 + 50666 = 2430652$$

= estimate of the variance of w_g

The variance of each estimator above has two main components, the first associated with the variation between sample plots, the second associated with the biomass table.

We shall now show that the common procedure consisting of the calculation of (i) the individual plot biomass per acre \hat{v}_h , (ii) average biomass per acre \bar{v} and (iii) the variance of \bar{v} as the variance of individual plot values \hat{v}_h divided by n , yields the first variance component. Using the individual plot biomass per acre as given by Cunia (1986e), we find

$$\sum \hat{v}_h = 12490431 \text{ for the first measurement}$$

$$= 14084491 \text{ for the second measurement}$$

$$\sum (\hat{v}_h)^2 = 196995482800 \text{ for the first measurement}$$

$$= 2295905587000 \text{ for the second measurement}$$

and the sum of the cross products between the first and the second measurement values \hat{v}_h , as equal to 2105734088000.

The reader can verify then that

$$S_{vv}^{11} = 5536998459, S_{vv}^{22} = 5254525850, S_{vv}^{12} = 5255342828$$

$$w_1 = \bar{v} \text{ of the first measurement}$$

$$= 12490431/118 = 105851.11$$

$$w_2 = \bar{v} \text{ of the second measurement}$$

$$= 14084491/118 = 119360.09$$

$$w_g = w_2 - w_1 = 13508.98$$

$$S_{w_1 w_1} = S_{vv}^{11}/n = 5536998459/118 = 46923716$$

$$S_{w_2 w_2} = S_{vv}^{22}/n = 5254525850/118 = 44529880$$

and

$$S_{w_g w_g} = (S_{vv}^{11} - 2S_{vv}^{12} + S_{vv}^{22})/118 = 2379986$$

Let us now see the percent of the total

error represented by the error of the biomass regression function. When estimating μ_1 and μ_2 the percentages are 18.58 and 20.37. The percentage becomes negligibly small, about 2 percent when estimating the net change μ_g . It appears that the effect of the error of the biomass regression may be ignored when the net change is estimated but the effect is sizable when the current averages are estimated. In this last case, it is also important to realize that we have a large sample of trees (from which the biomass regression function is estimated) and a small sample of plots. For illustration purposes let us consider a much more realistic sample of 88 trees (one fourth of our sample of 353 trees) and a sample of 472 plots (four times as many as our sample of 118 plots). Then, the variance of w_1 is expected to be about

$$(46923716/4) + (4) (10710584) = 11730929 + 42842336 = 54573265$$

and the percent of the total error due to the biomass regression is now about 78.50, an extremely large value.

Estimation of Growth Components μ_s , μ_m and μ_i

To estimate the average biomass growth on survivors and the biomass of the mortality and ingrowth trees, one should go to the original tree measurements and calculate the new plot values of $[s^s]$, $[s^m]$ and $[s^i]$. These new values are given by Cunia (1986e) and not repeated here. Using them, we calculate the following vectors of averages and the covariance matrices

$$[s^s] = \begin{bmatrix} 0 \\ 107.75 \\ 1346.0135 \end{bmatrix}, \quad [s^m] = \begin{bmatrix} 7.2033898 \\ 49.911017 \\ 470.95899 \end{bmatrix}$$

$$[s^i] = \begin{bmatrix} 36.864407 \\ 141.52542 \\ 546.95690 \end{bmatrix}$$

$$[S_{ss}^{ss}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3371.6046 & 36266.937 \\ 0 & 36266.937 & 437765.17 \end{bmatrix}$$

$$[S_{ss}^{mm}] = \begin{bmatrix} 148.52238 & 1066.5183 & 11480.546 \\ 1066.5183 & 9533.2480 & 123761.82 \\ 11480.546 & 123761.82 & 1840636.3 \end{bmatrix}$$

$$[S_{ss}^{ii}] = \begin{bmatrix} 1120.8533 & 4304.0120 & 16653.079 \\ 4304.0120 & 16595.353 & 64517.716 \\ 16653.079 & 64517.716 & 252214.66 \end{bmatrix}$$

Using the relationships $[z] = [S]$ and $[S_{zz}] = [S_{ss}]/n$, we find that

$$w_s = [b]'[z^s] = 14718.385 \text{ pounds} \\ = \text{estimate of the biomass growth per acre on survivor trees}$$

$$w_m = [b]'[z^m] = 4873.939 \text{ pounds} \\ = \text{estimate of the biomass per acre loss}$$

due to mortality

$$w_i = [b]'[z^i] = 3664.5330 \text{ pounds} \\ = \text{estimate of the biomass per acre growth due to ingrowth trees}$$

$$S_{w_s w_s} = [b]'[S_{zz}^{ss}][b] + [z^s]'[S_{bb}] [z^s] \\ = 439839 + 608965 = 1048804 \\ = \text{estimate of the variance of } w_s$$

$$S_{w_m w_m} = [b]'[S_{zz}^{mm}][b] + [z^m]'[S_{bb}] [z^m] \\ = 1996418 + 24567 = 2020984 \\ = \text{estimate of the variance of } w_m$$

and

$$S_{w_i w_i} = [b]'[S_{zz}^{ii}][b] + [z^i]'[S_{bb}] [z^i] \\ = 98330 + 767722 = 866052 \\ = \text{estimate of the variance of } w_i$$

It can be verified that

$$w_2 = w_i + w_s + w_m - w_m = 105851 + 14718 + 3665 - 4874 \\ = 119360$$

and

$$w_g = w_s + w_i - w_m = 14718 + 3665 - 4874 = 13509$$

The percent of the total error of the estimate of mortality, due to the error of the biomass regression function is negligible, about 1.22; this may be due to the fact that the variation of mortality from plot to plot is very large. On the other hand the percent is very high for survivor growth and ingrowth about 58.06 and 88.65 respectively. In the first case the pairs of values of the survivor trees are very highly correlated and thus, their differences have a small variance, and in the second case, the ingrowth trees are small in size, they are all estimated from the extreme side of the biomass regression function where the error is extremely large.

Estimation of the Average Biomass per Tree μ_t

We shall start with the calculation of the average biomass per tree for the first measurement. For notational convenience, the superscript 1, and later 2, will not be shown, and subscript h will be dropped. The set of original data has been summarized as the following basic statistics. Recall that $v_h = [b]'[s_h]$ is the biomass per acre of plot h, and $t_h = s_{h1}$ is the number of trees per acre of plot h.

$$E v = 12490431, \quad E t = \Sigma s_1 = 27890, \quad \Sigma s_2 = 169839.3, \\ \Sigma s_3 = 1285983.399, \quad S_{vv} = 5536998459, \\ S_{vt} = 5100342.7, \quad S_{tt} = 21272.932 (=S_{s_1 s_1})$$

The covariance matrix of [s] has already been given as $[S_{ss}^{11}]$ when μ_1 was estimated and

$$[S_{st}] = \begin{bmatrix} S_{s_1 t} \\ S_{s_2 t} \\ S_{s_3 t} \end{bmatrix} = \begin{bmatrix} S_{s_1 s_1} \\ S_{s_1 s_2} \\ S_{s_1 s_3} \end{bmatrix} = \begin{bmatrix} 21272.932 \\ 110452.37 \\ 602351.30 \end{bmatrix}$$

= sample covariance vector of [s] and t.

Using matrix notation, the reader can verify that

$$[z^t] = \begin{bmatrix} \Sigma s_1 \\ \Sigma s_2 \\ \Sigma s_3 \end{bmatrix} / \Sigma t = \begin{bmatrix} 1 \\ 6.0896128 \\ 46.109122 \end{bmatrix}$$

and

$$[S_{zz}^{tt}] = n \left([S_{ss}^{11}] - [z^t] [S_{st}]' - [S_{st}] [z^t]' + [z^t] [z^t]' \right) / (\Sigma t)^2$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 0 & .015573400 & .29380459 \\ 0 & .29380459 & 5.8076198 \end{bmatrix}$$

For example,

$$S_{z_2 z_3}^{tt} = (118)(4724541.6) - (6.0896128)(602351.30) - (46.109122)(110452.37) + (6.0896128)(46.109122)(21272.932) / (27890)^2 = .29380459$$

This yields immediately

$$w_t = [b]' [z^t] = 447.84622 \text{ pounds}$$

= estimate of the average biomass per tree at the first measurement

and

$$S_{w_t w_t} = [b]' [S_{zz}^{tt}] [b] + [z^t]' [S_{bb}] [z^t]$$

$$= 794.19269 + 191.72561 = 985.91831$$

= estimate of the variance of w_t

It can be easily verified that the first component of the variance of w_t is that given by the common procedure

$$S_{w_t w_t} = n (S_{vv} - 2w_t S_{vt} + w_t^2 S_{tt}) / (\Sigma t)^2$$

$$= (118) (5536998459 - (2)(447.84622)(5100342.7) + (447.84622)^2 (21272.932)) / (27890)^2 = 794.19269$$

since

$$w_t = \Sigma v / \Sigma t = 12490431 / 27890 = 447.84622$$

Similarly, to estimate the average biomass per tree at the second measurement time we start with the basic statistics

$$\Sigma t = \Sigma s_1 = 31390, \Sigma s_2 = 193364.3,$$

$$\Sigma s_3 = 1453780.747, S_{tt} = 24039.555$$

$$[S_{ss}] = [S_{ss}^{22}] \text{ as given before and}$$

$$[S_{st}] = \begin{bmatrix} 24039.555 \\ 123557.78 \\ 635935.23 \end{bmatrix}$$

Then the reader can verify that

$$[z^t] = \begin{bmatrix} 1 \\ 6.1600605 \\ 46.313499 \end{bmatrix},$$

$$[S_{zz}^{tt}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & .012967567 & .24205016 \\ 0 & .24205016 & 4.7494072 \end{bmatrix}$$

$$w_t = [b]' [z^t] = 448.69355$$

= estimate of the average biomass per tree at the second measurement

and

$$S_{w_t w_t} = [b]' [S_{zz}^{tt}] [b] + [z^t]' [S_{bb}] [z^t]$$

$$= 648.44837 + 161.00944 = 809.45780$$

= estimate of the variance of w_t

As a simple check of some of our calculations, we find

$$w_t (\Sigma t) / n = 105851.11 = w_1 = \text{for the first measurement}$$

$$= 119360.09 = w_2 = \text{for the second measurement}$$

Extension to More than One Species

The formulae above can be easily extended to more than one species. We start with the giant size vector [B] of the regression coefficients vectors $[b^1], [b^2], \dots, [b^p]$ of species 1, 2, ..., p respectively, that is,

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

and the covariance matrices $[S_{bb}^{11}], [S_{bb}^{22}], \dots, [S_{bb}^{pp}]$

arranged in the giant size covariance matrix

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

It is assumed that $[b^i]$ is statistically independent of $[b^j]$, for $i \neq j = 1, 2, \dots, p$. If not, one should calculate all the covariance matrices $[S_{bb}^{ij}]$ of $[b^i]$ and $[b^j]$ and complete the matrix $[S_{BB}]$ above. One way in which $[S_{bb}^{ij}] \neq 0$ may arise

is shown by Cunia (1986d).

For each species we calculate the plot vector $[s]$, say $[s^1]$, $[s^2]$, ..., $[s^p]$ of species 1, 2, ..., p respectively. After constructing the giant size vector

$$[s_h]' = [[s_h^1]' [s_h^2]' \dots [s_h^p]']$$

for each plot h, we calculate, by the usual formulae

$$[Z] = [\bar{S}] = \text{vector of the averages of } [s_h]$$

and

$$[S_{ZZ}] = (\text{covariance matrix of } [s_h])/n$$

Then,

$$w_1 = [B]'[Z] = \text{estimator of } \mu_1 \text{ (of all species)}$$

and

$$S_{w_1 w_1} = [B]'[S_{ZZ}][B] + [Z]'[S_{BB}][Z]$$

= estimator of the variance of w_1

The same approach can be applied to estimate the other parameters of interest, such as, for example, μ_2 , μ_g or μ_t .

Concluding Remarks

The common procedure for the calculation of the error of volume estimates from Continuous Forest Inventory data ignores the effect of the error of the tree biomass tables; it is implicitly assumed that this error is negligibly small. And this may indeed be true whenever (i) the number of CFI plots or points is relatively small, (ii) the sample of trees from which the biomass tables was constructed is relatively large and (iii) we are concerned with estimators of current volumes per unit area or per tree. Then, the error of the sample plots is much larger than the error of the biomass tables.

We have reconsidered the approach proposed by Cunia (1965, 1986a) to add the error of biomass tables to the error of sample plots, when the total error of biomass estimates is calculated. We have extended the applicability of this approach to CFI systems for estimates of (i) average biomass per unit area, (ii) average biomass per tree and (iii) average biomass per unit areas for growth components such as net change between measurements, mortality, ingrowth or survivor growth. We have also shown how to extend the applicability of the formulae from plots of fixed area to Bitterlich sample points.

We have then illustrated how to apply these formulae to an actual case of 118 permanent sample plots and a biomass regression function calculated from 353 sample trees. We have seen that, with this combination of sample sizes, the percent of the total error associated with the biomass tables is only about 20 percent, when the estimates of average biomass per acre at the

first and second measurement are calculated. The percent remains the same for the average biomass per tree. We felt, however, that these sample sizes are not representative of what is actually occurring in real life. Working with four times as many plots (472) and four times as few sample trees (88) we have shown that the expected part of the error of the biomass tables may rise to about 80 percent. This size of the error is indeed very high and can hardly be ignored.

The effect of the error of biomass regression function may be quite different when we consider estimates of growth components. It may be negligibly small for estimates of net change and mortality and considerably large for estimates of survivor growth and ingrowth. Sometimes this can be explained. The ingrowth is calculated from small trees, whose biomass regression estimates have a large error; they are all calculated from the extreme of the regression function where the error is at its maximum.

In the case of survivor growth, the error due to sample plots is small (relative to that of biomass tables) because the differences between the two measurements of the same trees have a very small variance; the correlation between the two measurements is extremely high. We are at a loss, however, to explain why the part of the error due to biomass regression function is small for net change and mortality; it is possible that the variation of the net change and mortality from plot to plot overshadows completely in size the error of biomass tables.

We have completely ignored the possible bias of the biomass tables calculated for one, and applied to another forest area. It is assumed that the two forest areas are sufficiently similar for the bias to be small. However, this may not be true. Because the bias is non-statistical in nature and can hardly be estimated, it is generally ignored. By properly designing (i) the sampling method for selecting sample trees for biomass regression function and (ii) the statistical procedure by which the regression function is calculated, one may be able to quantify the error component due to the difference between forest areas or stands, and, thus, take it into account when calculating the total error of the biomass estimates. More research is, however, needed in this area.

We have finally shown that the extension of the methodology from one to several species is straightforward. It requires the construction of (i) giant size vectors $[B]$ containing the individual species vectors $[b]$ and their covariance matrix $[S_{BB}]$ and (ii) giant size vectors of sample plot values $[s]$ corresponding to vector $[B]$ and their averages $[\bar{S}] = [Z]$ and covariance matrices $[S_{ZZ}] = [S_{SS}]/n$. Then, the formulae for the estimator w of μ and its variance are the same.

Acknowledgements

This paper is based on research funded by the Research Foundation of the State University of New York, the United States Department of Agriculture Forest Service and the Department of

Literature Cited

- Cochran, W.G. Sampling Techniques, 3rd Ed. John Wiles and Sons, New York, NY; 1977.
- Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.
- Cunia, T. On the error of forest inventory estimates: Continuous Forest Inventory with SPR. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986c.
- Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.
- Cunia, T. On the error of biomass estimates in forest inventories: Part 2: the error component from sample plots. Faculty of Forestry Miscellaneous Publication Number 9 (86-001), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986e.

ON THE ERROR OF FOREST INVENTORY ESTIMATES:

CONTINUOUS FOREST INVENTORY WITH SPR. (1)

Tiberius Cunia

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

In a previous paper, Cunia has shown how to take into account the error due to biomass tables when the error of forest biomass estimates is calculated in Continuous Forest Inventory (CFI). This approach is extended to the case of CFI systems using Sampling with Partial Replacement (SPR) methodology. The formulae to calculate the error of current biomass estimates and the estimates of growth components are given for the case of SPR on two successive measurements when (i) biomass tables are constructed from linear regression functions for which an estimate of the covariance matrix of the regression coefficients is given and (ii) the CFI plots or relascope sample points are selected by simple random sampling independently of the given biomass regression functions.

Introduction

Although it is known that the biomass regressions applied to forest inventory are not without error, it is a common procedure to ignore this error when inferences are made about the average biomass per unit area. Cunia (1986a) has proposed one approach to combine the error from inventory sample plots with the error from biomass regressions when (i) the regressions are linear and the vector [β] of m coefficients is estimated by the vector [b] calculated by least squares techniques, (ii) there are m statistics z₁, z₂, ..., z_m calculated from the sample plots data, so that the vector [μ_z] of their expected values multiplied by the vector [β] is equal to μ, the parameter one wishes to estimate and (iii) the vectors [b] and [z] are statistically independent. He gave the formulae for the estimators

$$w = [b]'[z] \text{ of the parameter of interest } \mu,$$

and

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

of the variance of w, where [S_{zz}] and [S_{bb}] are the known estimators of the covariance matrices of [z] and [b] respectively. Note that the left hand side of the expression of the variance

(1) Paper based on a set of lecture notes

"On the error of biomass estimates in forest inventory: Part 2: the error component from sample plots". Faculty of Forestry Miscellaneous Publication Number 9 (ESF 86-001). SUNY College of Environmental Science and Forestry, Syracuse, NY.

of w may be viewed as an error component due to sample plots, while the right hand side may be viewed as the error component due to biomass regressions.

In his paper, Cunia (1986a) considers the case where the sample plots are selected by simple random sampling and the parameter of interest μ is the mean biomass per unit area. This procedure was then extended, see Cunia (1986a), to the case where the sample plots were all permanent and part of a Continuous Forest Inventory (CFI) system and the parameters of interest are the mean biomass per acre at the first and second occasion (μ₁ and μ₂), the net change in the mean biomass per acre between the first and second occasion (μ_g = μ₂ - μ₁), the mean biomass per tree (μ_t) and the mean biomass growth per acre by growth component (μ_m for mortality, μ_i for in-growth and μ_s for growth on survivor trees.

It is known that CFI systems can be made more efficient when, at each measurement time, Sampling with Partial Replacement (SPR) methodology is being used and a part of the old sample plots is being replaced by new plots. A relationship between past and present measurement values is established from the data of the plots that are being remeasured, and this relationship is being used to update the values of the old plots that are not being remeasured and backdate the values of the new plots that have not been measured in the past.

It is the objective of the present paper to derive the formulae for the estimation of (i) current values and growth for CFI systems using SPR, and (ii) the error of these estimates that includes the error of biomass regressions in addition to the error of the CFI plots or points. For the general approach of how to combine the error of biomass regressions with that of the sample plots, the reader is sent back to Cunia (1986a,b). For the general methodology of SPR as applied to Forest Inventory the reader is sent to papers by Ware and Cunia (1962), Cunia (1965), Cunia and Chevrou (1969) and Newton, Cunia and Bickford (1974) among others.

We shall use the terminology and notation introduced by Cunia (1986a,b), in particular that of the cluster (plot) variables s₁, s₂, ..., defined on a "per unit area, acre or hectare" basis. For convenience, these variables will be defined in a later section. Because these variables can be defined on both sample plots of fixed area or relascope sample points, it is immaterial whether CFI sample units are plots or points; for convenience we shall use plots to denote them both.

Only CFI systems with SPR on two occasions will be considered here. To facilitate the derivation of the methodology that combines the error from sample plots and volume regressions, we shall present the SPR formulae in a more streamlined matrix form which may differ from that used in the above-mentioned SPR papers.

The Basic Methodology of SPR on Two Occasions

We shall start with a summary description of the basic methods of calculating the SPR estimators of μ_1 = average biomass per unit area at the first measurement time, μ_2 = average biomass per unit area at the second measurement time, and $\mu_g = \mu_2 - \mu_1$ = average change in biomass from the first to the second measurement time. We shall assume that there are (u + m + n) sample plots selected by simple random sampling, where u (for unmatched) plots were measured on the first but not on the second occasion, m (for matched) plots were measured on both occasions, and n (for new) plots were measured on the second but not on the first occasion.

To describe the methodology, we shall need additional notation. The biomass regression functions are usually defined in terms of dependent and independent variables y and x respectively. Because we work only with vector [b] and matrix [S_{bb}], we shall use x and y to denote different values. More specifically, we shall use variables x_1, x_2, \dots, x_p to denote measurements made on sample plots on the first occasion, and variables y_1, y_2, \dots, y_q to denote measurements made on sample plots on the second occasion. For convenience, we shall arrange these variables as vectors [x] and [y]. Except for x_1 and y_1 defined as biomass per unit area on the first and second occasion respectively, the other x and y variables (which we shall call here auxiliary) may or may not be paired with each other.

Let the vector of the averages of x_1, x_2, \dots, x_p , as calculated by the usual formulae from the u unmatched plots be denoted as

$$[\bar{x}_u]' = [\bar{x}_{1u} \bar{x}_{2u} \dots \bar{x}_{pu}]$$

where []' means transpose of []. Similarly, the vector of the averages of y_1, y_2, \dots, y_q , as calculated from the n new plots is denoted as

$$[\bar{y}_n]' = [\bar{y}_{1n} \bar{y}_{2n} \dots \bar{y}_{qn}]$$

and the vectors of averages of the m matched plots will be denoted as

$$[\bar{x}_m]' = [\bar{x}_{1m} \bar{x}_{2m} \dots \bar{x}_{pm}]$$

and

$$[\bar{y}_m]' = [\bar{y}_{1m} \bar{y}_{2m} \dots \bar{y}_{qm}]$$

The sample estimators of the variances and covariances of x and y variables are derived from the m matched plots only, by the usual formulae. Denoted as $S_{x_i x_j}, S_{x_i y_j},$ and $S_{y_i y_j}$ they are arranged

in the three covariance matrices [S_{xx}] of [x], [S_{xy}] of [x] and [y] and [S_{yy}] of [y]. Note that the order of [S_{xy}] is p by q, since the covariance terms of a given variable x_i are arranged in the i-th row, while the covariance terms of a given variable y_j are arranged in the j-th column.

The SPR estimators, say \hat{x}_1 of μ_1, \hat{y}_1 of $\mu_2,$

and \hat{g} of $\mu_g = (\mu_2 - \mu_1)$ are now defined as the 3 by 1 vector

$$[w] = \begin{bmatrix} \hat{x}_1 \\ \hat{y}_1 \\ \hat{g} \end{bmatrix} = [Q] + [A]'[P]$$

where

$$[Q] = \begin{bmatrix} \bar{x}_{1u} \\ \bar{y}_{1n} \\ \bar{y}_{1n} - \bar{x}_{1u} \end{bmatrix} = \text{vector of the averages of unmatched (and new) plots}$$

$$[P] = \begin{bmatrix} [\bar{x}_u] - [\bar{x}_m] \\ [\bar{y}_n] - [\bar{y}_m] \end{bmatrix} = \text{vector of the (p + q) differences between the averages of the matched and unmatched (and new) plots}$$

[A] = [G]⁻¹[H] = the (p + q) by 3 matrix of the SPR coefficients

$$[G] = \begin{bmatrix} \left(\frac{u+m}{um}\right) [S_{xx}] & \left(\frac{1}{m}\right) [S_{xy}] \\ \left(\frac{1}{m}\right) [S_{xy}]' & \left(\frac{m+n}{mn}\right) [S_{yy}] \end{bmatrix}$$

= covariance matrix of [P]

[G]⁻¹ = inverse of [G]

and

$$[H] = \begin{bmatrix} S_{x_1 x_1} / u & 0 & -S_{x_1 x_1} / u \\ \vdots & \vdots & \vdots \\ S_{x_1 x_p} / u & 0 & -S_{x_1 x_p} / u \\ 0 & S_{y_1 y_1} / n & S_{y_1 y_p} / n \\ \vdots & \vdots & \vdots \\ 0 & S_{y_1 y_q} / n & S_{y_1 y_q} / n \end{bmatrix}$$

= covariance matrix of [P] and [Q]

If we define, in addition

$$[K] = \begin{bmatrix} S_{x_1 x_1} / u & 0 & -S_{x_1 x_1} / u \\ 0 & S_{y_1 y_1} / n & S_{y_1 y_1} / n \\ -S_{x_1 x_1} / u & S_{y_1 y_1} / n & (S_{x_1 x_1} / u + S_{y_1 y_1} / n) \end{bmatrix}$$

then, the estimator of the covariance matrix of [w] is

$$[S_{ww}] = [K] + [A]'[H] = [K] - [H]'[G]^{-1}[H]$$

Various SPR estimators are generally obtained whenever different sets of auxiliary variables are being used. All these estimators ignore the effect of the biomass regression function used in the calculation of x_1 and y_1 . To take this effect into account, we may use the approach of the next section.

The New SPR Estimators

To take into account the error of the volume tables, we need plot variables s_1, s_2, \dots of the type defined by Cunia (1986 a,b). For more details the reader is referred to these papers. It suffices to say here that if the regression function of u = tree biomass on some independent variables v_1, v_2, \dots, v_r is of the form

$$\hat{u} = b_1 v_1 + b_2 v_2 + \dots + b_r v_r$$

then, the plot variables are defined as

$$s_i = \Sigma(v_{ik}/a_k), \quad i = 1, 2, \dots, r$$

where Σ is taken over all trees k in a plot or point and a_k is the factor that converts the measurement v_{ik} of tree k to a "per unit area" basis. For example, if $v_1 = 1$ and $v_2 = d$ = tree diameter, and the tree is selected from a plot of fixed area of "a" acres, then

$$s_1 = \Sigma(v_1/a) = \Sigma(1/a) = \text{number of trees per acre, and}$$

$$s_2 = \Sigma(v_2/a) = \Sigma(d/a) = \text{sum of diameters per acre}$$

We start with the same three samples of u unmatched, m matched and n new plots. Each plot is measured for the variables s_1, s_2, \dots, s_r on the first, on the second, or on both occasions depending on whether it is one of the u , one of the n or one of the m plots respectively. Using wherever necessary superscripts to denote the first (1) or the second (2) measurement, and subscript u, m and n to denote the sample from which the statistics are calculated, we define

(1) the vectors of sample averages

$$[\bar{s}_u]' = [\bar{s}_{1u} \quad \bar{s}_{2u} \quad \dots \quad \bar{s}_{ru}]$$

$$[\bar{s}_n]' = [\bar{s}_{1n} \quad \bar{s}_{2n} \quad \dots \quad \bar{s}_{rn}]$$

$$[\bar{s}_m^1]' = [\bar{s}_{1m}^1 \quad \bar{s}_{2m}^1 \quad \dots \quad \bar{s}_{rm}^1] \text{ and}$$

$$[\bar{s}_m^2]' = [\bar{s}_{1m}^2 \quad \bar{s}_{2m}^2 \quad \dots \quad \bar{s}_{rm}^2]$$

(2) the covariance matrices of $[s^1]$ and $[s^2]$ calculated from the data of the m matched plots alone,

$$[S_{ss}^{11}] = \text{sample covariance matrix of } [s^1], \text{ the vector of the first measurement values } s_1, s_2, \dots, s_r$$

$$[S_{ss}^{12}] = \text{sample covariance matrix of } [s^1], \text{ the vector of the first and } [s^2], \text{ the}$$

vector of the second measurement values s_1, s_2, \dots, s_r

and

$$[S_{ss}^{22}] = \text{sample covariance matrix of } [s^2], \text{ the vector of the second measurement values } s_1, s_2, \dots, s_r$$

Note (i) the use of superscripts 1 and 2 for the statistics of the m matched plots and the dropping of these superscripts when they are not needed, (ii) the covariance terms of $[S_{ss}^{12}]$ associated with s_i of the first measurement are all on the row i , while those associated with s_j of the second measurement are all on the column j , and (iii) $[S_{ss}^{21}] = [S_{ss}^{12}]'$.

Consider now the following vectors

$$[Q_1] = [\bar{s}_u], \quad [Q_2] = [\bar{s}_n], \quad [Q_g] = [Q_2] - [Q_1]$$

and

$$[P] = \begin{bmatrix} [\bar{s}_u] - [\bar{s}_m^1] \\ [\bar{s}_n] - [\bar{s}_m^2] \end{bmatrix}$$

The covariance matrices of these vectors are

$$[K_1] = [S_{ss}^{11}]/u = \text{estimator of the covariance matrix of } [Q_1]$$

$$[K_2] = [S_{ss}^{22}]/n = \text{estimator of the covariance matrix of } [Q_2]$$

$$[K_g] = [K_1] + [K_2] = \text{estimator of the covariance matrix of } [Q_g]$$

$$[G] = \begin{bmatrix} \left(\frac{u+m}{um}\right) [S_{ss}^{11}] & \left(\frac{1}{m}\right) [S_{ss}^{12}] \\ \left(\frac{1}{m}\right) [S_{ss}^{12}]' & \left(\frac{m+n}{mn}\right) [S_{ss}^{22}] \end{bmatrix}$$

= estimator of the covariance matrix of $[P]$

$$[H_1] = \begin{bmatrix} [K_1] \\ [0] \end{bmatrix} = \text{estimator of the covariance matrix of } [P] \text{ and } [Q_1]$$

$$[H_2] = \begin{bmatrix} [0] \\ [K_2] \end{bmatrix} = \text{estimator of the covariance matrix of } [P] \text{ and } [Q_2]$$

and

$$[H_g] = \begin{bmatrix} -[K_1] \\ [K_2] \end{bmatrix} = \text{estimator of the covariance matrix of } [P] \text{ and } [Q_g]$$

where

$[0]$ is the r by r matrix of zero.

We can define now the SPR estimators of the expected values of s_1, s_2, \dots, s_r denoted here as the mean vectors $[\mu_1^1]$ for the first measurement, $[\mu_2^2]$ for the second measurement, and $[\mu_2^g] = [\mu_2^2] - [\mu_2^1]$ for the net change from the first to the second

measurement as the statistics

$$[z^1] = [Q_1] + [A_1]'[P]$$

$$[z^2] = [Q_2] + [A_2]'[P]$$

and

$$[z^g] = [Q_g] + [A_g]'[P]$$

where

$$[A_1] = [G]^{-1}[H_1]$$

$$[A_2] = [G]^{-1}[H_2]$$

and

$$[A_g] = [G]^{-1}[H_g] = [A_2] - [A_1]$$

The covariance matrices of $[z^1]$, $[z^2]$ and $[z^g]$ are estimated by the formulae

$$[S_{zz}^{11}] = [K_1] + [A_1]'[H_1] = [K_1] - [H_1]'[G]^{-1}[H_1]$$

$$[S_{zz}^{22}] = [K_2] + [A_2]'[H_2] = [K_2] - [H_2]'[G]^{-1}[H_2]$$

and

$$[S_{zz}^{gg}] = [K_g] + [A_g]'[H_g] = [K_g] - [H_g]'[G]^{-1}[H_g]$$

Finally, the SPR estimators of μ_1 = average volume per unit area on the first occasion, μ_2 = average volume per unit area on the second occasion, and μ_g = average change in volume per unit area from the first to the second occasion are respectively defined as

$$w_1 = [b]'[z^1], w_2 = [b]'[z^2], \text{ and } w_g = [b]'[z^g]$$

Their variances are estimated by the formulae

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

$$S_{w_2 w_2} = [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2]$$

and

$$S_{w_g w_g} = [b]'[S_{zz}^{gg}][b] + [z^g]'[S_{bb}][z^g]$$

Note that the new SPR estimators are generally different than the old SPR estimators of the previous section. This is because different sets of auxiliary variables are being used. However, the difference between the old and new estimators is expected to be small and probably not significantly different than zero from a statistical point of view.

Illustrative Numerical Example

Consider the sample data from $u = 117$ temporary plots measured in 1960, $m = 118$ permanent plots measured on two occasions in 1960 and 1964, and $n = 92$ temporary plots measured only in 1964. These plots are one-tenth acre in size and have been selected from a forest area in Québec, Canada

by a method which can be assumed to be equivalent to simple random sampling.

To the trees of these plots we shall apply a biomass regression function used by Cunia (1986a,b). For the sample data and the calculation of the regression function

$$u = b_1 + b_2 d + b_3 d^2$$

where u = tree biomass (total above ground green weight in pounds) and d = tree diameter (inches), the reader is referred to Cunia (1986c). We need here only the statistics

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

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

Because (i) the independent variables of the biomass regression function are 1 , d and d^2 , and (ii) the plot size is one-tenth of an acre, the plot variables s_1 , s_2 , and s_3 are defined as

$$s_1 = \Sigma(1/.10 \text{ acres})$$

$$= \text{number of trees per acre}$$

$$s_2 = \Sigma(d/.10 \text{ acres})$$

$$= \text{sum of tree diameters per acre, and}$$

$$s_3 = \Sigma(d^2/.10 \text{ acres})$$

$$= \text{sum of squared tree diameters per acre}$$

The values of s_1 , s_2 , and s_3 as measured in 1960 and/or 1964 on the $(u+m+n) = 327$ plots are given in a set of lecture notes by Cunia (1986d). We shall now use these values to calculate three sets of SPR estimators.

First Set: The New SPR Estimators -

To calculate these estimators, we shall need the following summary statistics calculated from the basic plot values s_1 , s_2 , and s_3

$$[\bar{s}_u] = \begin{bmatrix} 214.78632 \\ 1338.3291 \\ 10722.382 \end{bmatrix} = \text{vector of the 1960 averages calculated from the data of the } u = 117 \text{ plots}$$

$$[\bar{s}_n] = \begin{bmatrix} 228.69565 \\ 1502.8380 \\ 12521.600 \end{bmatrix} = \text{vector of the 1964 averages calculated from the data of the } n = 92 \text{ plots}$$

$$[\bar{s}_m^1] = \begin{bmatrix} 236.35593 \\ 1439.3161 \\ 10898.164 \end{bmatrix} = \text{vector of the 1960 averages calculated from the data of the } m = 118 \text{ plots}$$

$$[\bar{s}_m^2] = \begin{bmatrix} 266.01695 \\ 1638.6805 \\ 12320.176 \end{bmatrix} = \text{vector of the 1964 averages calculated from the data of the } m = 118 \text{ plots}$$

$$[S_{ss}^{11}] = \begin{bmatrix} 21272.932 & 110452.37 & 602351.30 \\ 110452.37 & 659011.32 & 4724541.6 \\ 602351.30 & 4724541.6 & 48604076 \end{bmatrix}$$

= estimate of the covariance matrix of the first measurement values s_1 , s_2 , and s_3 as calculated from the $m = 118$ plots

$$[S_{ss}^{12}] = \begin{bmatrix} 21987.926 & 118295.97 & 659678.73 \\ 109292.82 & 670898.34 & 4832348.2 \\ 542577.80 & 4443890.0 & 46552860 \end{bmatrix}$$

= estimate of the covariance matrix of the first measurement values s_1 , s_2 , and s_3 with the second measurement values s_1 , s_2 , and s_3 as calculated from the $m = 118$ plots

$$[S_{ss}^{22}] = \begin{bmatrix} 24039.555 & 123557.78 & 635935.23 \\ 123557.78 & 718316.31 & 4802637.7 \\ 635935.23 & 4802637.7 & 47000205 \end{bmatrix}$$

= estimate of the covariance matrix of the second measurement values s_1 , s_2 , and s_3 as calculated from the $m = 118$ plots

Note that (i) we have used superscripts to denote the first or second measurement (do not interpret them as exponents or powers), and (ii) the covariance terms of $[S_{ss}^{12}]$ related to the first measurement variables are arranged in rows and those of the second in columns. For example, the covariance of the first measurement s_1 with the second measurement s_2 is 118295.97 while the covariance of s_1 of the second measurement with s_2 of the first measurement is 109292.82.

By using now the formulae of the new SPR estimators, one can easily calculate and write down the vectors

$$[Q_1], [Q_2], [Q_g], \text{ and } [P].$$

and their covariance matrices

$$[K_1], [K_2], [K_g], [G], [H_1], [H_2], \text{ and } [H_g]$$

These are not given here, but the interested reader can find their values listed in Cunia (1986d). The inverse of $[G]$ is also listed there as well as the calculation of the following vectors of SPR coefficients.

$$[A_1] = \begin{bmatrix} -.57582423 & .27162431 & 2.1942318 \\ -.016295496 & -.71115549 & -.60394450 \\ .00082660796 & .0038778696 & -.60368650 \\ -.052395398 & -1.4196676 & -6.1216962 \\ .075951873 & .58362944 & 1.2028896 \\ -.0031360528 & -.011748779 & .23558038 \end{bmatrix}$$

$$[A_2] = \begin{bmatrix} .58503065 & 1.0689709 & 1.9045323 \\ -.051364277 & .14848762 & .068781280 \\ .0017007641 & .0047393886 & .30871816 \\ -.62752621 & .39234747 & 2.7713606 \\ -.021883489 & -.81539047 & -.73733567 \\ .0011022853 & .0052754490 & -.67282885 \end{bmatrix}$$

and $[A_g] = [A_2] - [A_1]$ not shown here.

Consequently the SPR estimates of

$$[\mu_z^1] = \begin{bmatrix} \text{mean number of trees per acre in 1960} \\ \text{mean sum of diameters per acre in 1960} \\ \text{mean sum of squared diameters per acre in 1960} \end{bmatrix}$$

$[\mu_z^2]$ similarly defined for 1964 measurement,

and

$$[\mu_z^g] = [\mu_z^2] - [\mu_z^1]$$

are the following

$$[z^1] = \begin{bmatrix} 219.71326 \\ 1374.9418 \\ 10954.679 \end{bmatrix}, [z^2] = \begin{bmatrix} 247.57976 \\ 1561.1367 \\ 12280.514 \end{bmatrix},$$

$$\text{and } [z^g] = \begin{bmatrix} 27.866502 \\ 186.19492 \\ 1325.8344 \end{bmatrix}$$

and the estimates of their covariances are

$$[S_{zz}^{11}] = \begin{bmatrix} 65.995681 & 342.03116 & 1869.1504 \\ 342.03116 & 2039.9531 & 14673.089 \\ 1869.1504 & 14673.089 & 151545.20 \end{bmatrix}$$

$$[S_{zz}^{22}] = \begin{bmatrix} 75.556641 & 386.92007 & 1995.4171 \\ 386.92007 & 2243.7140 & 15044.211 \\ 1995.4171 & 15044.211 & 147808.27 \end{bmatrix}$$

and

$$[S_{zz}^{gg}] = \begin{bmatrix} 8.2793523 & 38.357716 & 219.65978 \\ 38.357716 & 209.87519 & 1597.4279 \\ 219.65978 & 1597.4279 & 17693.325 \end{bmatrix}$$

Consequently, the SPR estimators of μ_1 , μ_2 , and μ_g are calculated respectively as

$$w_1 = [b]'[z^1] = 108150.31 \text{ pounds}$$

$$w_2 = [b]'[z^2] = 120738.64 \text{ pounds}$$

$$w_g = [b]'[z^g] = 12588.336 \text{ pounds}$$

and their variances are estimated as

$$S_{w_1 w_1} = [b]'[S_{zz}^{11}][b] + [z^1]'[S_{bb}][z^1] = 17292284 + 11049512 = 28341796$$

$$S_{w_2 w_2} = [b]'[S_{zz}^{22}][b] + [z^2]'[S_{bb}][z^2] = 16553974 + 11720724 = 28274698$$

and

$$S_{w_g w_g} = [b]'[S_{zz}^{gg}][b] + [z^g]'[S_{bb}][z^g]$$

$$= 2078027 + 43969 = 2121996$$

Note that the part of the total variance due to the error of biomass regression function is about 38.98, 41.45 and 2.07 percent respectively for w_1 , w_2 , and w_g . It may be interesting to compare these percentages with those obtained by Cunia (1986b) when the CFI estimators were calculated from the $m = 118$ permanent sample plots alone. When estimating the current average biomass per acre (w_1 and w_2) the error component due to the biomass regression is about the same in absolute terms (11049512 versus 10710584 for w_1 and 11720724 versus 11393835 for w_2) but about twice as large in relative terms (38.98 versus 18.58 for w_1 and 41.45 versus 20.37 for w_2). This is because the part of the total variance of the SPR estimators due to the error of the sample plots is smaller in absolute terms (17292284 versus 46923716 for w_1 and 16553974 versus 44529880 for w_2) than that of the CFI estimators; the SPR estimators are based on data from 353 plots while the CFI estimators are based on the data of the 118 permanent plots only. On the other hand, the error component of the variance of w_g due to the error of the biomass tables remains about the same for the SPR and CFI estimators, (43969 versus 50666 for the absolute and 2.07 versus 2.08 percent for the relative terms).

Second Set: The Common (Old) SPR Estimators

We thought it would be interesting to calculate the values (estimators and errors) by the common SPR procedure and compare them to the values of the first set. If we define the plot variables as

x_1 = plot biomass per acre at the first, 1960 measurement time, and

y_1 = plot biomass per acre at the second, 1964 measurement time we can proceed as follows.

We start with the calculations of individual plot values x_1 and y_1 . The ordinary procedure is to calculate the biomass of each individual tree separately, sum them up by sample plots and divide these sums by .10 acres, the area of the plot. Because the individual plot variables s_1 , s_2 , and s_3 are available, it is easier to use the formulae $x_1 = [b]'[s^1]$ and $y_1 = [b]'[s^2]$. Cunia (1986d) lists the individual plot values x_1 and y_1 and calculates the following basic statistics that are needed for the calculation of the second set of the SPR estimators

$$\bar{x}_u = 106046.85 = [b]'[\bar{s}_u]$$

$$\bar{y}_n = 125267.64 = [b]'[\bar{s}_n]$$

$$\bar{x}_m = 105851.11 = [b]'[\bar{s}_m^1]$$

$$\bar{y}_m = 119360.09 = [b]'[\bar{s}_m^2]$$

$$s_{xx} = 5536998459, s_{xy} = 5255342828,$$

$$s_{yy} = 5254525850$$

Note that there is only one variable (x and y) measured on the plot on a given occasion, not $p > 1$ and $q > 1$ as implicitly assumed in our formulae. This would simplify the calculation of the SPR estimators. Also, the statistics S_{xx} , S_{xy} , and S_{yy} were calculated only from the data of the $m = 118$ permanent plots.

Using the appropriate formulae, the reader can verify that

$$[Q] = \begin{bmatrix} \bar{x}_u \\ \bar{y}_n \\ \bar{y}_n - \bar{x}_u \end{bmatrix} = \begin{bmatrix} 106046.85 \\ 125267.64 \\ 19220.790 \end{bmatrix}$$

$$[P] = \begin{bmatrix} \bar{x}_u - \bar{x}_m \\ \bar{y}_n - \bar{y}_m \end{bmatrix} = \begin{bmatrix} 195.73999 \\ 5907.5513 \end{bmatrix}$$

$$[K] = \begin{bmatrix} 47324773 & 0 & -47324773 \\ 0 & 57114411 & 57114411 \\ -47324773 & 57114411 & 104439185 \end{bmatrix}$$

$$[G] = \begin{bmatrix} 94248489 & 44536804 \\ 44536804 & 101644292 \end{bmatrix}$$

$$[G]^{-1} = (10^{-9}) \begin{bmatrix} 13.380774 & -5.8629647 \\ -5.8629647 & 12.407167 \end{bmatrix}$$

$$[H] = \begin{bmatrix} 47324773 & 0 & -47324773 \\ 0 & 57114411 & 57114411 \end{bmatrix}$$

$$[A] = \begin{bmatrix} -.63324208 & .33485978 & .96810186 \\ .27746347 & -.70862804 & -.98609152 \end{bmatrix}$$

= matrix of SPR coefficients

Consequently, the (common) SPR estimates are

$$[w] = \begin{bmatrix} w_1 \\ w_2 \\ w_g \end{bmatrix} = [Q] + [A]'[P] = \begin{bmatrix} 107562.03 \\ 121146.93 \\ 13584.900 \end{bmatrix}$$

and

$$[S_{ww}] = [K] + [A]'[H] =$$

$$= \begin{bmatrix} 17356735 & 15847163 & -1509572 \\ 15847163 & 16641538 & 794375 \\ -1509572 & 794375 & 2303947 \end{bmatrix}$$

Note that one can verify that

$$S_{w_g w_g} = S_{w_1 w_1} - 2S_{w_1 w_2} + S_{w_2 w_2}$$

$$= 17356735 - (2) (15847163) + 16641538 = 2303947$$

It may be interesting to compare these results to those obtained before as the first set. As one can easily verify the results are not identical. This is not surprising, since the two SPR procedures do not use the same auxiliary variables. The second set is based only on the biomass per acre variables of the first and second measurement, while the first set is based on plot variables s_1 , s_2 , and s_3 .

The differences between the two sets of measurements may look, at first sight, large. If we consider first the difference between the estimates w_1 and w_2 , we find the values

$$108150.31 - 107562.03 = 588.28 \text{ or about } .54 \text{ percent}$$

and

$$120738.64 - 121146.93 = -408.29 \text{ or about } .34 \text{ percent}$$

Obviously these differences are negligibly small, about 11.05 and 7.68 percent of the standard error of w_1 and w_2 respectively. The difference is, however, much larger for the estimate w_3 since

$$12588.336 - 13584.900 = -996.564 \text{ or about } 7.9 \text{ percent}$$

But this is still only about 68.41 percent of the standard error of w_3 of the second set of estimators.

The variances of the second set of SPR estimators are much smaller than those of the first set. This is not surprising since the error of the second set does not include the error of the biomass regression function. When this last error is taken out from the error of the first set, the estimates of the variances become extremely close. The slightly smaller values of the first set are due to the fact that the first set uses three auxiliary variables s_1 , s_2 , s_3 , while the second set uses only a single variable, the biomass per acre, that is nothing but a linear combination of s_1 , s_2 and s_3 , say x or y equal to $b_1s_1 + b_2s_2 + b_3s_3$.

To see what happens when the common (old) SPR estimators are calculated with s_1 , s_2 , and s_3 used as the auxiliary variables, we have calculated, as an example the following additional SPR estimator of μ_2 .

Third (Old) SPR Estimator of μ_2 : s_1 , s_2 , and s_3
Used as Auxiliary Variables

To calculate this estimator, we proceed as follows:

$$[Q] = [\bar{y}_n] = [125267.64]$$

$$[P] = \begin{bmatrix} [\bar{s}_u] - [\bar{s}_m] \\ \bar{y}_n - \bar{y}_m \end{bmatrix} = \begin{bmatrix} -21.569607 \\ -100.98704 \\ -175.78196 \\ 5907.5513 \end{bmatrix}$$

$$\{K\} = [S_{yy}/n] = [57114411]$$

$$[G] = \begin{bmatrix} \left(\frac{u+m}{um}\right) [S_{ss}^{11}] & \left(\frac{1}{m}\right) [S_{sy}^{12}] \\ \left(\frac{1}{m}\right) [S_{sy}^{12}]' & \left(\frac{m+n}{mn}\right) [S_{yy}^{11}] \end{bmatrix}$$

$$= \begin{bmatrix} 362.09902 & 1880.0743 & 10252.974 & 47859.527 \\ 1880.0743 & 11217.417 & 80419.186 & 390847.37 \\ 10252.974 & 80419.186 & 827318.40 & 4181841.0 \\ 47859.527 & 390847.37 & 4181841.0 & 101644292 \end{bmatrix}$$

$$[H] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ S_{yy}/n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 57114411 \end{bmatrix}$$

Note that to calculate $[G]$ we need the additional covariance terms of s_1 , s_2 , and s_3 (of the first measurement) with y (the biomass per acre of the second measurement). The values of these covariances divided by $m = 118$ are shown in $[G]$ as 47859.527, 390847.37, and 4181841.0.

The reader can verify that

$$[A] = [G]^{-1}[H] = \begin{bmatrix} 12.787013 \\ -6.5898007 \\ 4.0713794 \\ -.71009051 \end{bmatrix}$$

$$w_2 = [Q] + [A]'[P] = 120746.74$$

and

$$S_{w_2 w_2} = 16558010$$

Note that the newly obtained values (of the old SPR estimators) of 120746.74 and 16558010 are much closer to the corresponding values (of the new SPR estimators) of 120738.64 and 16553974.

Concluding Remarks

The approach used by Cunia (1986a) to combine the error of volume tables with that of sample plots when calculating volume and growth estimates in Continuous Forest Inventory (CFI) systems is extended to the case where the Sampling with Partial Replacement (SPR) is used to increase the efficiency. We have considered the simplest case, that of forest inventories on two successive occasions. By properly defining the cluster variables s_1 , s_2 , ..., one can apply the same formulae to clusters of both types, sample plots of fixed area or relascope sample points.

The new SPR estimators are slightly different (although, for our sample data not significantly different) than the classical ones where the error of the biomass tables is being ignored. This is due to the fact that the classical SPR method is applied to linear combinations of cluster variables s_1 , s_2 , ... (namely cluster biomass per unit area), while the new SPR approach is applied directly to

these variables. In this sense the new SPR estimators would normally be better; SPR regressions on linear combinations of auxiliary variables would not generally be better than regression on the variables themselves.

We have considered only one species. By defining the giant size vector [B] of regression coefficients

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

where $[b^i]$ is the vector of regression coefficients of species $i = 1, 2, \dots, c$, and its giant size covariance matrix $[S_{BB}]$ whose ij -th submatrix component is the covariance matrix $[S_{bb}^{ij}]$ of $[b^i]$ and $[b^j]$, one can extend the methodology from one to more than one species. This requires the definition of a giant size vector of cluster (plot or relascope point) values

$$[s]' = [[s^1]' [s^2]' \dots [s^c]']$$

where $[s^i]$ is the subvector of cluster values corresponding to species i , and the giant size covariance matrix $[S_{SS}]$ of $[s]$ whose ij -th submatrix component is the covariance matrix $[S_{ss}^{ij}]$ of $[s^i]$ and $[s^j]$. From here on, the procedure and the formulae introduced here apply immediately; the giant size vector $[Z]$ is defined as

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

where $[z^i] = [\bar{s}^i]$ = the sample average of $[s^i]$, and the giant size covariance matrix $[S_{ZZ}]$ contains as submatrix component $[S_{zz}^{ij}] = [S_{ss}^{ij}]/n$.

The approach can also be generalized to more than two measurements by using the SPR approach outlined by Cunia and Chevrou (1969) and Newton, Cunia and Bickford (1974). It suffices to define the vectors $[Q]$ and $[P]$ so as to (i) include the averages from all measurements on all auxiliary variables, and (ii) have the SPR estimators $w = [Q] + [A]'[P]$ unbiased.

Acknowledgements

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

Literature Cited

- Briggs, E.F.; Cunia, T. Effect of cluster sampling in biomass tables construction: linear regression models. *Canadian Journal of Forest Research*, 12: 255-263; 1982.
- Cunia, T. Continuous forest inventory, partial replacement of samples and multiple regression. *Forest Science*, 11: 480-502; 1965.

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

Cunia, T. On the error of forest inventory estimates: Continuous Forest Inventory without SPR. In: *Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates"*, May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.

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

Cunia, T. On the error of biomass estimates in forest inventories: Part 2: the error component from sample plots. *Faculty of Forestry Miscellaneous Publication Number 9 (86-001)*, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

Cunia, T.; Chevrou, R.B. Sampling with partial replacement on three or more occasions. *Forest Science* 15: 204-224; 1969.

Jacobs, M.W.; Cunia, T. Use of dummy variables to harmonize tree biomass tables. *Canadian Journal of Forest Research*, 10:483-490; 1980.

Newton, C.M.; Cunia, T.; Bickford, C.A. Multi-variate estimators for sampling with partial replacement on two occasions. *Forest Science*, 20:106-116; 1974.

Schreuder, H.T.; Swank, W.T. A comparison of several statistical models in forest biomass and surface area estimation. In: *Forest Biomass Studies*. H.E. Young (Ed.), University of Maine at Orono, Maine; 1971.

Ware, K.D.; Cunia, T. Continuous forest inventory with partial replacement of samples. *Forest Science Monograph* 3; 1962.