

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

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

RESEARCH PAPERS

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

Moderator: Eric H. Wharton

EVALUATING ERRORS OF TREE BIOMASS REGRESSIONS BY
SIMULATION

Tiberius Cunia

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

To construct tree biomass regression functions for use in forest biomass inventory, one would normally (i) select a sample of trees by some random procedure, (ii) measure the biomass of the sample trees, and (iii) calculate a biomass regression function by some statistical procedure. The sampling method is generally complex and its structure is seldom taken into account by the estimation procedure. To see the effect of sampling method and estimation procedure on the error of the inferences made when the tree biomass regressions are used, simulation techniques were applied to a population of 22,753 forest trees distributed over 927 one-fifth acre plots. Samples of trees were repeatedly selected by different methods and, for each sample, biomass regressions were calculated by various statistical procedures. Applied to the parent population, the biomass regressions generated estimates of the average biomass per unit area and their error. From the analysis of the differences between true and estimated average biomass, conclusions were drawn about the bias, precision, and estimated precision of combinations of sampling method and estimation procedure.

Introduction

The sampling designs used in forest biomass inventory consist generally of two phases. In the first phase, a relatively large sample of trees (usually in clusters defined as trees growing on plots of fixed area or Bitterlich sample points) is selected by some statistical sampling procedure. These trees are measured for species, diameter, height, etc. but not measured for biomass. In the second phase, a relatively small sample of trees is randomly selected and its trees are measured for biomass in addition to species, diameter, height, etc. A tree biomass regression function (on species, diameter, etc.) is calculated from the second phase sample trees and this regression function applied to the trees of the first phase yields estimates of the average biomass per unit area or total biomass in the entire forest area.

Because of the structure of the sampling design, the error of the biomass estimates has two main components; one component due to the error of the first phase and a second component due to the error of the tree biomass regression function of the second phase. This last component is usually ignored when the error of the biomass estimates is calculated; it is difficult to assess the error of biomass regression in meaningful terms and although

the methodology to combine the two components exists, it is largely unknown to the forest mensurationists.

The objectives of the present study are those of evaluating the error of biomass regression functions as applied to forest inventory data for the purpose of estimating the average biomass per unit area. For this, the error of the biomass regression should be expressed in a suitable form; it must be combined with the error of the first phase sample data when the error of the average biomass per unit area is being estimated. In our study we shall assume that, to combine the errors from the two phases we shall assume that the approach suggested by Cunia (1965, 1986) is being used. This approach requires that (i) the biomass regression functions be of the linear form

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

where y = tree biomass, $x_1 = 1$, and x_2, x_3, \dots, x_m are the independent variables defined in terms of diameter, height, species, etc, and (ii) valid estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix $[\sigma_{bb}]$ of $[b]$ exist and are known. Consequently, the error expression of the biomass regression function used in our study is $[S_{bb}]$.

Because the error of the regression is a function of (i) the sampling design by which the sample trees are selected, and (ii) the estimation procedure by which the regression is calculated, we shall consider both these factors when evaluating this error. We shall not consider, however, (i) the error of measurement of the biomass of the sample trees when determined by subsampling; we shall assume that the trees are measured without error, (ii) the error of the first phase sample; we shall apply the biomass regression to the entire tree population, and (iii) the error made when the trees selected from one are applied to another tree population.

To attain our objectives, we have used computer simulation techniques consisting of the following main steps. We have started with the construction of a real world type population of some 23,000 trees distributed in clusters (sample plots of fixed area) and measured, among other things, for diameter, height and biomass. This population was stored in the computer. Using a variety of sampling methods, sample trees were then selected by computer and their diameter, height, and biomass recorded. There were 100 repetitions of this entire simulated sampling process resulting in 100 samples of trees, hopefully independent, for any given sampling method and sample size. In the next step, biomass regression functions were calculated by a variety of estimation procedures, one set of regressions for each sample of trees generated by the previous step. Each tree biomass regression was then applied to the tree population to estimate the known value of the average biomass per unit area. Finally, the hundreds of thousands of estimated values were analyzed for possible bias, error and efficiency of each combination of sampling method and estimation procedure.

Construction of the Tree Population

The basic data used to construct our tree population consists of field measurements of diameter at breast height d , merchantable height h_m , and species, among other things, performed on 22,753 trees contained in 927 one-fifth acre plots selected by some random procedure from the New York State forest lands. Only the merchantable trees were included in the population (trees of diameter not less than 5 inches) and 260 of the 927 sample plots were empty, that is, did not contain trees of merchantable size. The total height h and the total above ground biomass y (green weight) was not measured in the field but was generated for each population tree by a Monte Carlo procedure described in detail in a series of papers by Cunia and Michelakackis (1983b, 1984a, b) and Cunia, Michelakackis and Lee (1984). Because the procedure by which h and y was generated is an important factor to consider when interpreting the results of the analysis of the simulated sample data, the method to generate h and y is summarized below as follows.

The total height of each tree was generated first by a formula of the form $h = \hat{h} + q$ where (i) h is the simulated total tree height, (ii) \hat{h} is the regression estimate of the conditional mean of height for given species, diameter d , merchantable height h_m and geographical area, and (iii) q is a random variable expressing the difference between actual tree height and average height as estimated by the regression function. The regression function of total height on diameter and merchantable height was estimated from actual sample data of some 1,600 trees of twenty species selected from three states, New York, Michigan and West Virginia. The least squares method was used and the regression function was assumed to be of the linear form $\hat{y} = b_1 + b_2 d + b_3 h_m$. Separate regression functions by species and state were calculated and most of the time either d or h_m was sufficient to adequately define the regression relationship. This means that for most regressions, either b_2 or b_3 was made equal to zero. The regression coefficients b_1 , b_2 , and b_3 were stored in the computer and applied to our tree population data to calculate the values \hat{h} .

The random variable q was generated by a Monte Carlo technique according to a probability distribution estimated from the same sample trees. It was assumed that the shape but not the scale of the probability distribution of q was the same for all species. If (i) e is a standardized random variable with mean zero and variance one, and (ii) S_{hh} is the conditional variance of total tree height for given diameter, merchantable height, species and state, the random variable q was defined as being equal to $e\sqrt{S_{hh}}$. The variance S_{hh} was estimated separately for each sample regression but the probability distribution of $e = q/\sqrt{S_{hh}}$ was estimated from the pooled data of all regressions. As the sample distribution was found to be irregular in shape, a two-stage graphical procedure was used to smooth it out.

Using the total height h as generated above, the biomass of each population tree was generated

by a similar procedure. The formula used was of the form $y = \hat{y} + q$ where (i) y is the value of the tree biomass obtained by this simulation process, (ii) \hat{y} is the regression estimate of the conditional mean of biomass for given species, diameter d , total height h , geographical region and cluster (plot) and (iii) q is a random variable expressing the difference between actual tree biomass and the conditional mean biomass as estimated by the regression function. The regression function of y on d and h was estimated from actual sample tree data from Finland, New York, Michigan, Ohio, and West Virginia. The weighted least squares method was applied to a regression function of the linear form $\hat{y} = b_1 + b_2 d^2 h$ under the assumption that, within each species and state, the conditional variance of y is proportional to $(d^2 h)^2$. The regression functions were calculated separately by species and state and the effect of site was estimated by regression with dummy variables techniques. To take the cluster effect into account, we have estimated the probability distribution of pairs of regression coefficients b_1 and b_2 calculated within each individual cluster (plot) from their own tree data.

To estimate the probability distribution of q , we have assumed that $q = e(d^2 h)\sqrt{S_{uu|v}}$, where (i) e is a standardized random variable with mean zero and variance one, (ii) $S_{uu|v}$ is the estimate of the conditional variance of the transformed variable $u = y/d^2 h$ for given diameter and height, and (iii) the conditional standard deviation of y given diameter d and height h is estimated by $d^2 h\sqrt{S_{uu|v}}$. The values of $S_{uu|v}$ were calculated separately by species and state and the probability distribution of $e = q/(d^2 h)\sqrt{S_{uu|v}}$ was estimated from the pooled data of all regressions. A graphical two-stage procedure was also used to smooth out the irregular shape of the sample probability distribution of e .

To generate the biomass of a given tree, the Monte Carlo technique proceeded as follows. For the trees of a given plot a set of regression coefficients b_1 and b_2 was generated for each species according to the bivariate probability distribution of the pairs of plot values b_1 and b_2 . Using the tree diameter, height, species and the geographical region of the plot, the regression value \hat{y} was then calculated by the corresponding regression function. Finally, a value $q = e(d^2 h)\sqrt{S_{uu|v}}$ was added to \hat{y} , where (i) d and h are the tree diameter and height, (ii) $\sqrt{S_{uu|v}}$ is the stored value of the conditional standard deviation of $u = y/d^2 h$ for the given species and region, and (iii) e is a random variable with mean zero and variance one generated by the computer according to the probability distribution of e also stored in the computer.

Sampling Procedures

Although easily done by computer, it is relatively difficult in real life to select trees individually by simple random sampling. The usual procedure is to select the sample trees in clusters. Ordinarily defined as groups of trees contained in plots of fixed area, or trees counted by relascope from randomly selected points in the forest, the clusters may be selected by simple or stratified

random sampling, in one or two-phases and all the trees, or a randomly selected part of the trees of the sample clusters are measured for biomass, diameter, height and other attributes of interest. In our simulation study we have sampled trees by three main sampling procedures, denoted here, for convenience, as (i) two-stage random sampling, (ii) two-stage stratified sampling, and (iii) two-phase, two-stage random sampling. Within each sampling method we have defined several variants by changing the tree subsampling procedure of the second stage. Here is a more detailed description of our three sampling methods.

(1) Two-stage random sampling. In the first stage, m plots (clusters) are selected from the 667 non-empty plots of our population by simple random sampling without replacement. In the second stage, sample trees from the sample plots of the first stage are selected by one of seven subsampling procedures. In the first procedure, a fixed percentage p of trees is selected by simple random sampling without replacement. If a fixed number r of trees is selected, subsampling methods 2 and 3 are obtained when the selection is made without or with replacement respectively. The last four subsampling methods are obtained when the tree selection is made with replacement and with probability proportional to a measure of tree size; more specifically, proportional to tree height h , tree diameter d , tree basal area (d^2) and approximate tree volume (d^2h) respectively for methods 4, 5, 6, and 7.

The size of the sample is controlled by the number m of sample plots and the percentage p , or the fixed number r of trees subsampled within the plot. We have used the values $m = 1, 2, 5, 10, 15, 20, 30,$ and 50 , the values $p = .05, .10, .15, .30, .40, .60$ and 1.00 and the values $r = 1, 2, 5, 10, 15, 20,$ and 30 . In a second simulation study we have also used $p = .10, .20, \dots, 1.00$. Because we did not want samples that were too small or too large, we have used only those combinations of m with p or r for which the expected sample size fell between 10 and 500.

Note that the only sampling method resulting in selection of trees with equal probability from the entire population of trees is subsampling method 1. Selecting a fixed number of trees from each randomly selected plot (as in subsampling methods 2 and 3) results in a larger probability of selection for the big trees. It is known that the larger trees require more living space than the smaller trees and, thus, the big trees have the tendency to be contained in plots with a small, rather than a large number of trees. For the same reason, selecting fixed number of trees with probability proportional to $h, d, d^2,$ and d^2h (subsampling methods 4, 5, 6, and 7) results in tree selection with probability higher than $h, d, d^2,$ and d^2h respectively.

To reduce the amount of simulation work required, we have selected the sample plots and the trees within the sample plots in a nested fashion. To better describe this procedure, let us consider, as an illustrative example, the first subsampling method. In the first simulation run, the run that

generates the first sample of all sample sizes, 50 plots are selected by simple random sampling without replacement. Five percent of the trees selected from the first plot constitutes the first sample of size $m = 1, p = .05$. By adding five percent more trees from the same plot, we obtain the first sample of size $m = 1, p = .10$. But if we add instead five percent of the trees selected from the second plot, we obtain the first sample of size $m = 2, p = .05$. Continuing this way (in two directions, adding trees and plots), we finish with the selection of the first sample of the last sample size of $m = 50, p = 1.00$. Of course, some of these samples will be discarded; those that result in combinations of m and p for which the expected sample size is too small or too large.

Multiplying p by the number of trees in a given sample plot does not generally result in an integer number of sample trees. To decide whether an additional tree corresponding to the fractional part f (from .01 to .99) is to be selected, a random number R generated by the computer is being used. If $R \leq f$, an additional sample tree selected at random from the remaining unselected trees of the plot is added to the sample. Otherwise, no additional sample tree is selected.

When the subsampling of trees is done with replacement, the same tree may be selected in the sample more than once. This may present a problem when the sample plot has very few trees and the fixed number r is very large. For example, if $r = 20$ and the sample plot has one tree, the same tree will appear twenty times in the sample. To improve the sample, we have instructed the computer to generate a new biomass value for every tree already in the sample that happens to be selected again. The procedure to generate the new biomass is identical to that used when the tree population was constructed.

(2) Two-stage stratified sampling. With this method, the m sample clusters are selected by stratified random sampling. The tree population was first divided into three geographical regions (strata) and a two-stage random sample selected from each stratum separately. The same seven subsampling methods of tree selection from sample plots were used here as well. By combining now samples of various sizes from various strata, we obtain two-stage stratified samples of different sizes and allocations.

(3) Two-phase, two-stage random sampling. In the first phase, the trees from a relatively small two-stage random sample are measured for biomass y , diameter d and height h . The tree data are used to estimate the regression function of biomass on diameter and height, say $\hat{y} = r_1(d, h)$. To reduce the computer simulation work, the samples already selected under the two-stage random sampling method of (1) above were used as the first phase sample of the two-phase, two-stage random sampling method. In the second phase, a relatively large, two-stage random sample of trees is selected and every tree is measured for diameter and height but not for biomass. Only the first tree subsampling method is used with $m = 50, 100, 150, 200, 300,$ and 400 and $p = 2.93, 5.86, 8.79,$ and 11.72 .

The percentage p was selected so as to yield, on the average 1, 2, 3, and 4 sample trees per plot respectively. The data from this sample are used to estimate the regression function of tree height h on tree diameter d , say $\hat{h} = r_2(d)$. The two regression functions are then combined to obtain a regression function of biomass on diameter alone, say $\hat{y} = r(d)$.

To obtain a two-phase two-stage random sample, we must pair a sample of the first with a sample of the second phase. The sample trees of the first phase are selected by seven subsampling methods and there is a variety of sample sizes (values m and p or r) used. The sample trees of the second phase are selected by one subsampling method and there are 24 sample sizes, the product of six values m by four values p . For each simulation run, we obtain one two-phase, two-stage random sample for each combination of subsampling method and sample size of the first phase with a sample size of the second phase. As there are 100 simulation runs, there are 100 samples generated by the same two-phase two-stage random sampling procedure with the same sample size.

Estimation Procedures for the Biomass Regression Functions

Consider first the two-stage random sampling method. For each of the 100 samples generated for the same sample size (m and p or r) and tree subsampling procedure, we have calculated thirty-two biomass regression functions, the combinations of four least squares estimation approaches and eight regression equations.

The first estimation approach is that of the ordinary least squares (OLS) method applied to individual tree data. The fact that several sample trees may have been selected from the same cluster is not taken into account. Because the conditional variance of the tree biomass for given diameter or given diameter and height is not homogeneous, it is more reasonable to use the ordinary weighted least squares (OWLS) method. The variance is assumed to be proportional to d^4 for the regression of biomass on diameter alone, or proportional to d^4h^2 for the regression of biomass on diameter and height. With the OWLS approach the cluster effect is also ignored.

The third and fourth estimation approaches are known here as the modified least squares (MLS) and modified weighted least squares (MWLS) methods. They take into account the cluster effect when the regression function is calculated. This is accomplished by applying the weighted least squares method to the cluster (in our case the plot) not the individual tree values. If the tree variables are denoted by y, x_1, x_2, \dots and if Σ means summation over the trees of a given plot, then the plot variables are defined as $\Sigma y, \Sigma x_1, \Sigma x_2, \dots$. In the MLS approach it is assumed that the conditional variance of Σy is proportional to the number of trees in the plot, while in the MWLS approach the conditional variance is assumed to be proportional to Σd^4 if only d is used in the regression or proportional to Σd^4h^2 if both d and h are used.

The basic justification and methodology of the modified least and weighted least squares is given in Cunia (1979, 1981) and Briggs and Cunia (1982).

For each least squares estimation approach we have used eight regression equations, the first three of biomass on diameter alone, the remaining five of biomass on diameter and height. The equations are all linear, and they contain terms of the form d, d^2, h, dh and d^2h . The significance of the addition of some of these terms was sometimes tested statistically.

Consider now the two-stage stratified sampling method. Using the conclusions reached from the analysis of the eight regression equations applied to the seven subsampling procedures of the two-stage random sampling method, we have selected two equations only, the best equation of biomass on diameter alone and the best equation of biomass on diameter and height. As the four least squares estimation approaches were again used, the total number of combinations of approach and equation is eight.

The trees were selected by stratified sampling and the stratification itself may be used when the regression function is estimated. We have defined nine procedures to take the stratification effect into account. These procedures range from one single regression function calculated for all strata with stratification effect completely ignored to three, separately and independently calculated regressions, one for each stratum. The procedures are not described here, and the interested reader can refer to Arabatzis and Cunia (1986) for a more detailed description. The total number of regression functions calculated for each sample is, thus, equal to 72, the product of two regression equations by four least squares estimation approaches and nine ways to take stratification into account.

With the two-phase, two-stage random sampling methods the estimation procedure is more complex. Twenty regression functions were calculated for each first phase sample, the combinations of four least squares estimation approaches (OLS, OWLS, MLS, and MWLS) and five linear regression equations of the form $\hat{y} = r_1(d, h) = a_1x_1 + a_2x_2 + \dots$, where $x_1 = 1$ and the other variables x are defined in terms of diameter d and height h (d, d^2, h, dh and d^2h). Because the conditional variance of the tree height for given diameter is approximately homogeneous, only two least squares estimation approaches (OLS and MLS) were used with the samples of the second phase. They are combined with four linear regression equations of height on diameter of the form $\hat{h} = r_2(d) = c_1x_1 + c_2x_2 + \dots$ where $x_1 = 1$ and the other variables x are of the form d^1 and d^2 . There is a total of eight regression functions for each second phase sample.

A method to combine a regression function $\hat{y} = r_1(d, h)$ of the first phase with a regression function $\hat{h} = r_2(d)$ of the second phase so as to obtain a regression function $\hat{y} = r(d)$ of biomass on diameter alone has been described by Cunia (1982), Cunia and Michelakackis (1983a), and Michelakackis and Cunia (1985, 1986). This method defines the regression function as

$$\hat{y} = r_1(d, \hat{h}) = b_1 + b_2 d + b_3 d^2 + b_4 d^3 + b_5 d^4$$

where the regression coefficients b_1, b_2, \dots, b_5 are functions of the coefficients a and c of $r_1(d, h)$ and $r_2(d)$ above and the specific independent variables included in the regression being dependent on the specific regressions $r_1(d, h)$ and $r_2(d)$ used. A method is also described to estimate the covariance matrix of b_1, b_2, \dots, b_5 when the regression coefficients a and c and their covariance matrices are given.

To calculate the 160 regression functions of the form $\hat{y} = r(d)$ for each two-phase, two-stage random sample (the 160 combinations of twenty regression functions $\hat{y} = r_1(d, h)$ of the first with eight regression functions $\hat{h} = r_2(d)$ of the second phase) would have been prohibitively expensive in terms of computer time and analysis work. Instead, we have selected the four best regressions of the first phase (one for each of OLS, OWLS, MLS, and MWLS approaches) and the two best regressions of the second phase (one for OLS and one for MLS). Pairing the OLS regression of the second with the OLS and OWLS regressions of the first phase and, similarly, the MLS of the second with the MLS and MWLS regressions of the first phase, results in four regressions of the form $\hat{y} = r(d)$ for each two-phase, two-stage random sample.

Application of the Biomass Regression Functions

The sampling is done for the purpose of deriving an estimate z of the mean biomass per acre μ of our tree population. The value of μ is obviously known. But by analyzing the probability behavior of z , we can make inferences about the bias and precision of a given estimation procedure applied to a given sampling method of a given sample size.

There is one estimate z for each estimation procedure and each simulated sample. If the regression function is denoted in matrix notation as

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

where $x_1 = 1$ and, if the error of this regression function is expressed as the covariance matrix $[S_{bb}]$ of $[b]$, it can be shown that

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

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

where the elements of $[\mu_x]$ are the known expected values of the elements of $[x]$ expressed on a "per acre" basis.

For example, if the regression function is

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

then

$$z = b_1 \mu_{x1} + b_2 \mu_{x2} + b_3 \mu_{x3}$$

where

$$\mu_{x1} = \text{mean number of trees/acre}$$

$$\mu_{x2} = \text{mean (sum of } d\text{)}/\text{acre}$$

$$\mu_{x3} = \text{mean (sum of } d^2\text{)}/\text{acre}$$

The formulae above are easy to prove. For each tree i of the population, the estimate of its biomass y_i is $\hat{y}_i = [b]'[x_i]$, where $[x_i]$ is the vector of the variables x of the tree i . Adding the estimated biomass of all 22753 trees yields an estimate $\hat{Y} = \sum \hat{y}_i$ of the total biomass $Y = \sum y_i$ of our population. Dividing now \hat{Y} by the total forest area of $A = 927/5 = 185.4$ acres, we obtain an estimate $z = \hat{Y}/A$ of the mean biomass per acre $\mu = Y/A$. Symbolically this can all be written as

$$z = \hat{Y}/A = \sum \hat{y}_i / A \\ = \sum (b_1 x_{1i} + b_2 x_{2i} + \dots) / A \\ = b_1 (\sum x_{1i} / A) + b_2 (\sum x_{2i} / A) + \dots \\ = b_1 \mu_{x1} + b_2 \mu_{x2} + \dots \\ = [b]'[\mu_x]$$

The formula of the estimator V of the variance of z follows immediately from the fact that (i) μ_x is known without error and (ii) if $[a]$ is a vector of fixed values and $[v]$ is a vector (of the same order) or random variables with covariance matrix estimated by $[S_{vv}]$, then the variance of $[z] = [a]'[v]$ is estimated by

$$S_{zz} = \sum \sum a_i a_j S_{v_i v_j} = [a]'[S_{vv}][a]$$

Analysis Procedure

For each simulated sample and each regression estimation procedure we have calculated the pair of random variables z and V , estimators of μ and variance of z respectively. The pair of confidence intervals ($z \pm t\sqrt{V}$), where $t = 2.0$ for the 95 percent and $t = 2.6$ for the 99 percent confidence level, were also calculated and whether μ fell below, within or above these intervals was recorded. The 100 simulation runs giving rise to the 100 samples may be viewed as 100 repetitions of a random experiment, each experiment generating two random variables z and V , two random intervals ($z \pm t\sqrt{V}$) and two, trinomially distributed random variables that show whether the confidence intervals include the parameter μ or fall below or above it. For all practical purposes, these 100 repetitions may be viewed as being statistically independent, and we may also assume that from one to the next simulation run, the probability distribution of the random variables and intervals remains the same.

To analyze the probability behavior of z , we have calculated, for each set of 100 simulated samples and for each regression model the following statistics, where Σ stands for summation over the sample values from 1 to 100.

$$\bar{z} = \Sigma z / 100 = \text{average of the 100 sample values } z \\ = \text{estimator of } \mu$$

$$(\bar{z} - \mu) = \text{estimator of the bias of } z$$

$$\bar{V} = \Sigma V / 100 = \text{average of the 100 sample values } V \\ = \text{estimator of the variance of } z \text{ under the} \\ \text{basic assumptions of the regression model} \\ \text{used}$$

$S_{zz} = \Sigma(z-\bar{z})^2/99$ = estimator of the variance of z with no assumptions made other than those of the statistical independence of the 100 simulation runs and no change in the probability distribution of z from one to the next run

$t = (\bar{z}-\mu)/\sqrt{S_{zz}/100}$ = sample value of t that is used to test the null hypothesis that the bias of z is equal to zero

\bar{V}/S_{zz} = ratio of the two estimates of the variance of z ,

and finally, the number of times μ is found to fall below, within or above the 95 and 99 percent confidence limits defined above.

Note that under the assumptions of (i) statistical independence of the various simulation runs and (ii) unchanging probability distribution of z from one to the next run, the statistics \bar{z} and S_{zz} are unbiased estimators of the unknown mean μ_{zz} and variance σ_{zz}^2 of the random variable z . As z is used as an estimator of the mean biomass per acre μ , the statistic $(z-\mu)$ can be used as an estimator of the true bias of z , say $(\mu_z - \mu)$. The null hypothesis that this bias is equal to zero can be tested by the t -statistic with 99 degrees of freedom. When $(\mu_z - \mu)$ is small, the statistic S_{zz} can be used as a measure of the efficiency of the sampling method and estimation procedure.

Note also that in real life only the estimator V of the variance of z is available. This estimator is valid, however, only when the basic assumptions of the regression model are strictly satisfied by the tree population we sample and the tree selection procedure we use. As these assumptions are seldom if ever satisfied, we can make inferences about the goodness of V as an estimator of the variance of z by comparing \bar{V} , the average of the 100 sample values V with the unbiased estimator S_{zz} . This comparison can be made by means of sample differences $(V-S_{zz})$, the average difference $(\bar{V}-S_{zz})$ or the sample ratio \bar{V}/S_{zz} . We have preferred using the ratio.

The analysis of the bias $(\bar{z}-\mu)$ of z , the precision S_{zz} of z , and the validity of the estimator V has been made in two main steps. In the first step we have compared the various regression models within a given sampling method, tree subsampling procedure and sample size. In this way we were able to identify and, thus, eliminate from further consideration, the least squares estimation approach and the regression equation that may yield consistently poor results, that is, estimators with large bias and/or poor precision.

In the second step we have compared the various sampling methods, subsampling procedures and sample sizes. More specifically, we have made inferences about the effect on bias, precision and estimated precision of z of the factors (i) sample size, that is number of clusters and number of sample trees per cluster, (ii) fixed percentage p or number r of trees selected from sample clusters, (iii) subsampling trees with or without replacement, (iv) subsampling trees with equal

or unequal probability and (v) the overall sampling method, the two-stage random or stratified sampling or two-phase, two-stage random sampling.

To do this analysis, we have constructed hundreds of detailed and summary tables and graphs showing how $(z-\mu)$, S_{zz} , \bar{V} , \bar{V}/S_{zz} , etc. vary with the characteristics of the sampling method and estimation procedure.

Main Conclusions Drawn From the Study

The detailed analysis and the conclusions reached as the result of this analysis are contained in a series of papers by Cunia (1985), Cunia and Gillespie (1985), Michelakackis and Cunia (1985, 1986), Gillespie and Cunia (1986) and Arabatzis and Cunia (1986). We shall not repeat them here. However, it may be of interest to give a summary view of these conclusions, some of which came as a big surprise.

When the regression model is suitably selected, the bias of z is generally small even when it is significantly different from zero. One regression function that generates a large bias is $\bar{y} = \beta d^2 h$; a function that should usually be avoided. The bias does not seem to be affected by (i) the sample size (number of clusters or number of sample trees per cluster), (ii) the selection of a fixed percentage p or a fixed number of trees from a sample cluster, (iii) whether the sampling is done with or without replacement, or (iv) the least squares regression approach or equation (if properly selected). It seems, however, to be somewhat affected by the probability of tree selection when the clusters are subsampled; it seems to increase from small, non-significantly different from zero when the subsampling is done with equal probability (methods 1, 2, and 3), to slightly larger values when the subsampling is done with probability proportional to h and d (methods 4 and 5) and the bias becomes significantly different from zero when the probability is proportional to tree basal area (method 6) or approximate tree volume (method 7). The bias is also affected by stratification when the allocation of sample trees to strata is poor and the stratification effect is not properly taken into account. The bias does not seem to be affected, however, by the two-phase sampling method considered in this study.

As expected, the precision of z as measured by S_{zz} is affected by (i) the overall sample size; the larger the sample, the smaller the error, (ii) the number of trees per cluster for given overall sample size; the smaller the number of trees, the smaller the error, (iii) the least squares estimation approach; the weighted least squares are slightly better than the least squares approaches, and similarly, the ordinary least or weighted least squares are also slightly better than the corresponding modified approaches, and finally, (iv) the regression functions of biomass on d and h are much better than those on diameter alone.

The fact that the least squares is almost as good as the weighted least squares approach came as a surprise. We did not expect the least squares

method to be that robust. Also as a surprise came the fact that no increase in precision is obtained when the probability of selection moves from equal to proportional to a measure of tree size. On the contrary, for the least squares method (OLS and MLS) the precision seems to decrease dramatically as we move from equal to probability proportional to d (or h), to basal area (d^2), and finally to approximate volume (d^2h).

Not completely unexpected, the precision of z does not seem to be affected by (i) sampling with or without replacement, (ii) selection of a fixed number or a fixed percentage of trees from a sample cluster, and (iii) the form of the regression function when this form is not poorly selected (as for example the form $\hat{y} = \beta d^2h$) or when the independent variables are not the same (when both d and h are used we have, as expected, a much better estimator than when only d is used).

The precision of z can also be estimated by V . As expected, the precision is grossly overestimated (error grossly underestimated) by the ordinary least and weighted least squares regressions based on the individual tree, not on the cluster data. This overestimation of precision increases with the size of the subsample; the larger the number of trees selected from the same cluster, the greater the overestimation. On the other hand, there seems to be a slight underestimation of the precision (overestimation of the error) when the modified least or weighted least squares methods are used. It is possible, however, that this underestimation is due to the fact that with large samples the effect of the finite population correction factor (for the sample clusters that may go as high as 5 percent) has been ignored.

Finally, the two-phase, two-stage sampling procedures we have used, yields unbiased (or at most with negligible bias) estimates and the method to calculate the precision of the estimates is correct, whenever the precision of the first and second phase regressions are properly evaluated.

The conclusions above hold true for the population as constructed here. We feel that our population resembles real world tree populations. The relationship between diameter and height as it appears in real world sample plots is preserved because we have worked with real world plots. The relationship between diameter, height and biomass was, on the other hand, generated by the computer; but this relationship is expected to imitate very closely what happens in real life. Finally, the cluster effect we have added to our population may not be equal in size but it is definitely similar in structure to what we expect in the real world.

Acknowledgments

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

Literature Cited

- Arabatzis, A. A.; Cunia, T. Error of biomass regressions: sample trees selected by stratified sampling. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates," May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Briggs, E. F.; Cunia, T. Effect of cluster sampling in biomass tables construction: linear regression models. *Can. Journal of Forest Research*. 12:255-263; 1982.
- Cunia, T. Some theory on the reliability of volume estimates in a forest inventory sample. *Forest Science*. 11:115-128; 1965.
- Cunia, T. On sampling trees for biomass tables construction: some statistical comments. In: Forest resource inventory workshop proceedings, Vol. 2, W. E. Frayer (Ed.), Colorado State University, Fort Collins, CO; 1979.
- Cunia, T. Cluster sampling and tree biomass tables construction. In: Interdivisional Proceedings, 17th IUFRO World Congress, September 6-12, 1981, Kyoto, Japan; 1981
- Cunia, T. On the error of tree volume tables and its effect on the precision of forest inventory estimates. In: Statistics in theory and practice: essays in honor of Bertil Matern. B. Ranney (Ed.). Swedish University of Agricultural Sciences, Section of Biometry, S-90183, Umea, Sweden; 1982.
- Cunia, T. Use of simulation techniques in tree biomass tables construction. In: Proceedings, The 1985 Symposium on system analysis in forest resources. December 8-11, 1985, University of Georgia, Athens, GA; 1985.
- Cunia, T. On the error of biomass estimates in forest inventory; Part 1: its major components. Faculty of Forestry Miscellaneous Publication Number 8 (ESF 85-004), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Cunia, T.; Gillespie, A. J. Cluster sampling and construction of biomass tables: results of a simulation study. In: Proceedings, Third southern biomass energy research conference, March 12-14, 1985, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL; 1985.
- Cunia T.; Michelakackis, J. On the error of tree biomass tables constructed by a two-phase sampling design. *Can. Journal of Forest Research*. 13: 303-313; 1983a.
- Cunia, T.; Michelakackis, J. A method to construct a forest biomass population model. In: Proceedings, Renewable resource inventories for monitoring changes and trends. J. F. Bell and T. Atterbury (Eds.), Oregon State University, Corvallis, OR; 1983b.

- Cunia, T.; Michelakackis, J. A Monte Carlo technique for generating total height of forest trees. Faculty of Forestry Miscellaneous Publication Number 4 (ESF 84-018), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1984a.
- Cunia, T.; Michelakackis, J. Constructing forest biomass populations for simulated sampling. Faculty of Forestry Miscellaneous Publication Number 5 (ESF 84-019), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1984b.
- Cunia, T.; Michelakackis, J.; Lee, S. Generating total tree heights by a Monte Carlo technique. In: Proceedings, 1983 Southern forest biomass workshop, June 15-17, 1983, Charleston, SC, R. F. Daniels and P. H. Dunham (Eds.), USDA Forest Service, Southeastern Forest Experiment Station, Asheville, NC; 1984.
- Gillespie, A. J.; Cunia, T. Error of biomass regressions: sample trees selected by cluster sampling. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates," May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Michelakackis, J.; Cunia, T. Construction of biomass tables by double sampling: preliminary results of a simulation study. In: Proceedings, Use of auxiliary information in natural resource inventories, October 1-2, 1985, Blacksburg, VA. R. G. Oderwald, H. E. Burkhart and T. E. Burk (Eds.), Society of American Foresters Publication No. SAF 86-01; 1985.
- Michelakackis, J.; Cunia, T. Error of biomass regressions: sample trees selected by double sampling. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates," May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.

ESTIMATION OF TREE BIOMASS TABLES BY CLUSTER

SAMPLING: RESULTS OF A SIMULATION STUDY

Andrew J. Gillespie and Tiberius Cunia

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

Tree biomass regressions are generally constructed (i) from samples of trees selected by cluster rather than simple random sampling and (ii) by least squares techniques that ignore the cluster effect. The results of a simulation study are reported whereby (i) samples of trees were selected by cluster sampling from a known tree population, (ii) the biomass regression functions were calculated by ordinary least squares methods and by methods modified to take the cluster effect into account and (iii) the estimates of the average biomass per acre were compared to the known true value of the tree population. These results show, among other things, that (i) the estimates of the average biomass per acre based on the ordinary and modified regression techniques are about the same but the estimates of the precision are grossly overstated by the ordinary least squares, and (ii) the bias is not significantly different than zero for all suitably selected models.

Introduction

A common sampling design for taking the biomass inventory of a given forest area is the two phase or double sampling method. Phase one consists of a large sample of trees measured for diameter at breast height, species, and possibly height, along with other variables. These trees are not measured for biomass. The phase two sample is a smaller sample of trees measured for biomass as well as species, diameter, height, and possibly other variables measured in the phase one sample. This sample is used to construct a regression function of biomass on several predictor variables, ordinarily species, diameter, and possibly height. This regression function is then applied to the data of the phase one sample to estimate the mean biomass per tree or mean biomass per unit area.

Simple random sampling of trees is seldom a practical method of sampling a forest. It is more efficient to apply the method of cluster sampling. Cluster sampling involves dividing the forest into overlapping or non-overlapping clusters defined as plots of fixed or variable area. A subset of these clusters is then selected and some or all of the trees from each sample cluster are selected and measured for the variables of interest. The advantage of cluster sampling is the reduced average sampling cost per tree that

is associated with the reduced movement of sampling crews over the forest. The disadvantage is the corresponding reduced amount of information per tree; trees growing close together will tend to be more similar than trees growing farther apart.

One common method of estimating biomass functions from the phase two sample is the method of ordinary least squares (OLS) or weighted least squares (OWLS). These methods assume that the sample trees are selected independently of each other. This assumption is clearly violated in cluster sampling; trees in a given cluster are not selected independently of each other. Cunia (1979) suggested three modifications of the least squares method for cluster sampling. The first modification, using ratio estimator models, has been applied by Kotimaki and Cunia (1981) to two cluster samples of trees. The second modification uses linear regression models, and has been applied to the same two samples by Briggs and Cunia (1982). In both cases, the researchers found that the intracluster correlations had little effect on the point estimates of mean biomass. The estimates of the error of the point estimates, however, were much larger under the modified procedures. The assumptions of the modified models were better satisfied by the actual sampling conditions than were the assumptions of the ordinary models. Therefore, it was assumed that the unmodified procedures underestimated the error and that the modified procedures were better at estimating the error.

The objectives of the present study are to verify and extend the results of Kotimaki and Cunia (1981) and Briggs and Cunia (1982) by examining the accuracy and precision of various biomass function estimation procedures as applied to cluster sampling. We shall only consider the error of the phase two sample; the error of the first phase sample will be ignored. To accomplish this, we shall apply simulation techniques in which a given population of trees is grouped into clusters and repeatedly sampled. All population parameters (including μ = mean biomass per acre) are known exactly. The biomass equation from the phase two sample will be applied to the population stand table, rather than a phase one estimate. The resulting estimation of mean biomass per acre may then be compared directly to the known parameter μ . In the earlier studies listed above, samples and estimates came from natural populations, for which the precise parameters were unknown. Estimates could not be compared to a fixed standard, but rather were analyzed intuitively by how well their inherent assumptions were satisfied.

The simulated population used in this study consists of 22,753 trees from 667 non-empty and 260 empty permanent one-fifth acre sample plots selected from the State of New York. The trees were measured in the field for their species, diameter at breast height d and merchantable height. The total height h and total biomass y (green weight above ground) were simulated for each tree by Monte Carlo techniques described in

detail by Cunia and Michelakackis (1983, 1984a,b) and Cunia, Michelakackis, and Lee (1984). In generating total height and biomass, the techniques took into account the effects of species, diameter, merchantable height, site quality, geographical region, and the intracluster correlation effect, as well as the random effect of the probability distributions of total height and biomass around their respective regression functions.

Sampling Procedures

Our sampling procedure is the following: m clusters ($m = 10, 15, 20, 30, \text{ or } 50$) are chosen at random (without replacement) from the population of 667 non-empty clusters, and p percent ($p = 10, 20, \dots, 100$) of the trees from each cluster are chosen (at random without replacement) and measured for biomass (y), diameter (d), and height (h). The average cluster size is about 34 trees. The study considers only those combinations of m and p which yield an average sample size $(m)(p)$ (34) of at least 10 trees but less than 600 trees. There are 36 such combinations which, for convenience, will be known here as sampling methods. For each of these 36 sampling methods, 100 independent samples of trees were generated by computer from the known population of 22,753 trees.^{1/} This yields a study set of $36 \times 100 = 3600$ different randomly selected cluster samples of trees.

Estimation Models

The main population parameter of interest is μ = average total biomass per acre, defined as the total biomass of all trees in the population divided by the total forest area occupied by the simulated population.^{2/} It is known that μ is equal to 115.549 thousands of pounds per acre.

Two sets of models are initially used to estimate μ . The first set consists of two versions of each of seven ratio estimators, for a total of 14 ratio estimator models. The second set includes four versions of each of five linear functions, for a total of 20 least squares regression models. Initially, we applied each of these 34 estimation procedures to each of the 100 samples consisting of 50 clusters with 30 percent subsampling. After analyzing the results, we selected two ratio estimators and eight least squares regression models for further analysis in application to the remaining 3500 samples of trees. The original 34 models are briefly described below; for more complete descriptions, the reader is referred to Gillespie (1985).

^{1/}This work was done by Sueh Fang Hsu, Graduate student, SUNY-CESF.

^{2/}Total area = (927 plots) \times (1/5 acre per plot) = 185.4 acres.

Ratio Estimators

Ratio estimator models assume that the relationship between tree biomass y and a highly correlated variable x is adequately described by the equation $y = Rx$, where for this study R is defined as μ_y/μ_x , the ratio of the arithmetic means of y and x , and x is defined as either d^2 or d^2h . Note that R is also the ratio of the means per acre or totals over the entire forest of y and x . The purpose of sampling is to estimate R , which is assumed unknown. Models 1 through 7 consist of seven estimation procedures with $x = d^2$; models 8 through 14 consist of the same estimation procedures with $x = d^2h$. These procedures are used to calculate r = the point estimate of R , and $S_{r,r}$ = the estimate of $\sigma_{r,r}$, the variance of r .

Letting first $x = d^2$ and taking the summation over all trees of a given sample, we define the following seven models:

Model 1.

The true regression model of y on x is assumed to be $y = Rx$, with the sample trees selected by simple random sampling and the conditional variance of y given x is constant (homoscedastic):

$$r_1 = \Sigma xy / \Sigma x^2$$

and

$$S_{r_1 r_1} = (\Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2) / (n-1) \Sigma x^2$$

Model 2.

Identical to model 1, but with the conditional variance of y given x assumed to be proportional to x :

$$r_2 = \Sigma y / \Sigma x$$

and

$$S_{r_2 r_2} = (\Sigma (y^2/x) - (\Sigma y)^2 / \Sigma x) / (n-1) \Sigma x$$

Model 3.

Identical to model 2, with the conditional variance of y given x assumed to be proportional to x^2 :

$$r_3 = \Sigma (y/x) / n$$

and

$$S_{r_3 r_3} = (\Sigma (y/x)^2 - (\Sigma (y/x))^2 / n) / (n-1) n$$

Model 4.

Ratio-of-means model which assumes only that the sample trees were collected by simple random sampling:

$$r_4 = \bar{y} / \bar{x}$$

where \bar{y} and \bar{x} are the arithmetic sample means of y and x respectively. This estimator is known to be biased (Cochran, 1977). The variance of r_4 is estimated by

$$S_{r_4 r_4} = \left(S_{yy} - 2(r_4)S_{xy} + (r_4)^2 S_{xx} \right) / n\bar{x}^2$$

where S_{xx} , S_{yy} , and S_{xy} are the usual sample variances and covariance of x and y .

Model 5.

Mean-of-ratios estimator, which also assumes only that the sample trees were collected in a simple random sample:

$$r_5 = \Sigma (y/x) / n = r_3$$

$$S_{r_5 r_5} = (\Sigma (y/x)^2 - (\Sigma (y/x))^2 / n) / n(n-1)$$

This estimator is also known to be biased (Cochran, 1977).

Model 6.

Modified ratio-of-means model, where the trees are assumed collected by a simple random sample of clusters. The modification utilizes the cluster variables

$$u_h = \text{cluster biomass} = \Sigma y \text{ in cluster } h, \\ h = 1, 2, \dots, m,$$

and m = number of clusters in the sample,

and

$$v_h = \Sigma x \text{ in cluster } h$$

with Σ taken over the trees of cluster h . The modified ratio-of-means estimate of R is calculated by

$$r_6 = \bar{u} / \bar{v} = ((\Sigma y) / m) / ((\Sigma x) / m) = r_4 = r_2$$

but the variance has a different formula:

$$S_{r_6 r_6} = \left(S_{uu} - 2(r_6)S_{uv} + (r_6)^2 S_{vv} \right) / m\bar{v}^2$$

where \bar{u} , \bar{v} , S_{uu} , S_{vv} , and S_{uv} are the usual sample means, variances, and covariance of u and v .

Model 7.

Modified mean-of-ratios model, where the trees are assumed collected by a simple random sample of clusters. The modification utilizes the cluster variables

$$w_h = \Sigma (y/x) = \text{sum of ratios } (y/x) \text{ in cluster } h.$$

$$n_h = \text{number of sample trees in cluster } h, \text{ and}$$

the modified mean-of-ratios estimates of R and

σ_{rr} are calculated by

$$r_7 = \Sigma w / \Sigma n = \left(\Sigma (y/x) \right) / n = r_5 = r_3$$

$$S_{r_7 r_7} = m \left(S_{ww} - 2(r_7)S_{wn} + (r_7)^2 S_{nn} \right) / n^2$$

where S_{ww} , S_{nn} , and S_{wn} are the usual sample variances and covariance of w and n . Although $r_2 = r_4 = r_6$, the variances and assumptions of each model are distinct; therefore these are all different models. Similarly, while $r_3 = r_5 = r_7$ and $S_{r_3 r_3} = S_{r_5 r_5} = S_{r_7 r_7}$, the assumptions of these three models are also distinct, hence they represent three different models.

Models 8 - 14 consist of the same seven estimators with $x = d^2 h$. Once r and S_{rr} have been estimated, we calculate

$z_i = r_i \mu_x$ = the point estimate of mean biomass per acre, and

$V_i = S_{r_i r_i} (\mu_x)^2$ = the estimate of the variance of z_i ,

where μ_x = the known average "sum of x " per acre

Regression Estimators

The study considers 20 procedures (models) for estimating linear regressions equations of y as a function of d or d and h . These models consist of 20 combinations of five linear regression functions with four least squares estimation approaches. The five linear functions considered are:

1. $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$
2. $\hat{y} = \beta_1 + \beta_3 d^2$
3. $\hat{y} = \beta_1 + \beta_4 d^2 h$
4. $\hat{y} = \beta_4 d^2 h$
5. $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2 + \beta_4 d^2 h + \beta_5 d h + \beta_6 h$

with function 5 retaining only those terms for which the estimated coefficients are statistically discernible from 0. These functions are expressed in the standard mathematical notation $\hat{y} = [\beta]'[x]$, where $[\beta]$ = the vector of regression coefficients, $[\beta]'$ denotes the transposed vector $[\beta]$, and $[x]$ denotes the vector of predictor variables corresponding to $[\beta]$.

We consider four approaches to calculate $[b]$, the point estimate of the vector $[\beta]$ of regression coefficients. The first approach is that of ordinary least squares (OLS), applied to the individual trees of the sample with the conditional variance of y given $[x]$ assumed to be constant. The second approach is that of ordinary weighted least squares (OWLS), also applied to the individual trees of the sample, but with the conditional variance of y given $[x]$ assumed to be proportional to (i) d^4 for models which use d alone or (ii) $d^4 h^2$ for models which use

both d and h as independent variables. These first two approaches are both well documented in statistical literature as, for example, by Draper and Smith (1981), and by Neter, Wasserman, and Kutner (1983).

Approaches 3 and 4 are modified regression techniques which utilize cluster variables, rather than individual tree values. Since most forest inventories utilize cluster samples (rather than simple random samples of trees), modified techniques may be more accurate and precise than standard techniques for which the assumptions are not satisfied. These modifications are briefly described below. For more detailed descriptions of the rationale and method of modification for cluster sampling, the reader is referred to Cunia (1979, 1981) and Briggs and Cunia (1982). Modification for approaches 3 and 4 consists of creating m new cluster variables u_h and v_h by summing y_i and x_i over each cluster h ($i = 1, 2, \dots, n_h$ and $h = 1, 2, \dots, m$, where n_h = number of trees in cluster h):

$$u_h = \sum y_i \text{ in cluster } h, \text{ and}$$

$$v_h = \sum x_i \text{ in cluster } h.$$

The new sample size will now be the number of non-empty clusters in the sample. The new model is now assumed to be $\hat{u} = [\beta]'[v]$. Comparison of approaches 1 and 2 versus approaches 3 and 4 will illustrate the error made by treating a cluster sample as if it had been collected by simple random sampling.

We assume that the variables y are normally distributed and selected independently within any given cluster, that the conditional variance of y given [x] is equal to $a_i^2 \sigma^2$, where a_i^2 is known for each tree and σ^2 is an unknown constant, and that the other assumptions of least squares methods are satisfied. Since u is defined as the summation of several random variables y in a cluster, it is reasonable to further assume that the conditional variance of u given [v] is proportional to $(\sum a_i^2) \sigma^2$. Approach 3 assumes that the variables u and v of the various clusters are statistically independent, and that $a_i^2 = 1$ so that the conditional variance of u given [v] is proportional to $\sum (1)^2 = n_h$, the number of sample trees in cluster h. This is the modified least squares (MLS) approach. Approach 4 makes the same assumptions except that a_i^2 is assumed to be d^4 (functions 1,2) or $d^4 h^2$ (functions 3,4,5). The conditional variance of u given [v] is assumed to be proportional to $\sum d^4$ or $\sum d^4 h^2$. This is the modified weighted least squares (MWLS) approach.

To summarize, regression models 15-19 consist of the 5 linear functions estimated by approach 1; models 20-24 are the 5 linear functions estimated by approach 2; models 25-29 are the 5 linear functions estimated by approach 3; and models 30-34 are the 5 linear functions estimated by approach 4.

Once we have obtained the estimate [b] of $[\beta]$ and $[S_{bb}]$ of $[\sigma_{bb}]$ = the covariance matrix of

[b], the estimators z_i and V_i for regression model i are given by:

$z_i = [b_i]'[\mu_i]$ = the point estimate of mean biomass per acre, and

$V_i = [\mu_i]'[S_{bb}][\mu_i]$ = the estimate of the variance of z_i ,

where $[\mu_i]$ = a column vector of "means per acre" of each of the respective predictor variables x included in model i. These means are all known exactly for the simulated population.

We point out that, in terms of calculations (but not assumptions), (i) models 3 and 5 are identical, (ii) models 8 and 18 are identical, and (iii) models 10, 12, and 23 are identical. We therefore consider 30 rather than 34 distinct estimation procedures.

Analysis Procedure

For a given estimation procedure i ($i = 1, 2, \dots, 34$) and each sample j from a given sampling method ($j = 1, 2, \dots, 100$), we calculate the sample statistic z_{ij} and sample variance V_{ij} which estimate the population mean μ and variance σ_{zz} respectively. We also note whether μ is above, within, or below the 95 percent and 99 percent confidence intervals of z_{ij} . Finally, we summarize all 100 estimates z_{ij} and V_{ij} for each model-sampling method combination with the statistics:

1. $\bar{z}_i = \sum z_{ij} / 100$
2. $S_{zz} = (\sum z_{ij}^2 - (\sum z_{ij})^2 / 100) / 99$
3. $\bar{V}_i = \sum V_{ij} / 100$
4. $(\bar{z}_i - \mu)$
5. $t_i = (\bar{z}_i - \mu) / \sqrt{S_{zz} / 100}$
6. The number of times (out of 100 trials) that μ was above, within, or below the 95 and 99 percent confidence intervals of z_{ij} .

\bar{V}_i is an average estimate of the variance of z_{ij} under the assumptions of the model i. If the model assumptions are strictly satisfied, then each V_{ij} and their average \bar{V}_i are unbiased estimators of the variance σ_{zz} of z_{ij} . Because the 100 estimates z_{ij} are generated by the same random process (sampling method and estimation procedure), they may be considered to be 100 statistically independent random variables. In this case, unbiased estimates of μ and σ_{zz} are given by the usual statistics z and S_{zz} .

The analysis focuses on the variation of these summary statistics between models for a given sampling method, and within each given model over varying sampling procedures.

For the initial analysis, we apply the 34 models to 100 samples from the 50 cluster, 30 percent subsampling method. The results of this analysis are given in Cunia and Gillespie (1985), and are briefly summarized below.

The bias of the ratio-of-means estimator (models 2, 4, 6, 9, 11, and 13) is small and generally not significantly different than zero for our sample data. This is true for both $x = d^2$ and $x = d^2h$. The other ratio estimators (models 1, 3, 5, 7, 8, 10, 12, 14) have statistically significant bias that may vary from 2 to 7 percent of the mean μ . The estimate S_{zz} of the variance of z is very high for models 1 and 8, and more reasonable for the other models. The estimate V of the variance of z is (i) much lower than S_{zz} for models which ignore the cluster effect, and (ii) slightly higher than S_{zz} for models which allow for the cluster effect. Hence the only acceptable ratio estimator models are models 6 and 13, ratio-of-means estimators which account for the effects of cluster sampling. The adjustment for cluster sampling does not affect the accuracy of the estimation procedure. It does affect the estimated precision of the point estimate; the average estimated precision \bar{V} for modified models 6 and 13 is approximately three times as large as \bar{V} of the corresponding unmodified models 4 and 11. This agrees with the findings of Kotimaki and Cunia (1981), and is illustrated by the behavior of the confidence interval; only the confidence intervals for models 6 and 13 included μ an acceptable number of times out of 100 replications.

For the linear regression models 15 to 34, this first analysis indicates that (i) regression models 18, 23, 28, and 33 (with the equation $y = \beta_4 d^2 h$) are very erratic, with large bias and low precision; (ii) the bias of the other 16 regression estimators is generally small and not statistically different from zero; (iii) the models which ignore the effect of cluster sampling tend to grossly overestimate the precision of the point estimators, with the estimate V much lower than the corresponding S_{zz} , and (iv) modified models are much better at estimating the precision of z_{ij} , as illustrated by the relatively small difference between \bar{V} and S_{zz} , as well as the degree of confidence interval reliability. This agrees so far with the findings of Briggs and Cunia (1982).

At this point, the field of models is narrowed down to two ratio estimators (models 6 and 13) and 8 regression estimators (models 15, 17, 20, 22, 25, 27, 30, and 32) for complete analysis over the remaining 35 sampling procedures. These models are selected because they all yielded generally unbiased estimates of μ but different estimates of the error. Part two of the analysis will look at the behavior of these models over varying sampling methods.

Analysis of estimated bias ($\bar{z} - \mu$).

The size of the estimated bias ($\bar{z} - \mu$) varies within and between the 36 sampling methods. A set of tables and graphs given by Gillespie (1985) shows this variation. Following are the main conclusions that we have derived from the analysis of these tables and graphs.

Within the models using d alone, none of the estimates of bias are significantly different from zero at the 95 percent confidence level. The conclusions are the same for the models using both d and h except for the 10 (and sometimes 15) cluster samples, for which the bias is significant. There seems to be no reason for this inconsistency. We have concluded that, in general, the bias of z is small and, for our data, not significantly different than zero.

It is of interest to mention also that (i) for models using d alone, the absolute value of the bias seems to remain approximately constant over samples of any size (except for the 10 or 15 cluster samples, where it is larger), and (ii) there is a tendency for the bias of the models with d and h to decrease when, for a fixed number of sample trees, the number of sample clusters increases. This last phenomenon is assumed to be due to random effects.

Analysis of $\sqrt{S_{zz}}$ the estimate of $\sqrt{\sigma_{zz}}$.

The statistic $\sqrt{S_{zz}}$ is defined as the estimate of the standard deviation (error) of z . For a given model-sampling method i , it is calculated from the 100 estimates z_{ij} , and is thus independent of any model assumptions concerning variance estimation. Table 1 lists $\sqrt{S_{zz}}$ for all 360 model-sampling method combinations.

As expected, $\sqrt{S_{zz}}$ decreases as average sample size increases and as the number of clusters increases for a fixed sample size. As percent subsampling increases for a fixed number of clusters, $\sqrt{S_{zz}}$ decreases until approximately 30 to 40 percent of trees on a cluster have been sampled, then remains approximately constant as percent subsampling increases. Models using d alone have consistently higher $\sqrt{S_{zz}}$ (lower precision) than similar models using both d and h ; this is due to the additional information contributed by h in the estimation of biomass. Models using d alone show greater improvement in precision due to increases in the number of clusters sampled, with the difference in the precision estimates of these two groups becoming negligible when 50 clusters are sampled.

The OLS, OWLS, and ratio estimator models which use the same independent variables have roughly equal standard errors for any sampling method. The same relationship exists between MLS and MWLS models. This shows that, with respect to the point estimation of μ , unweighted regression procedures are very robust; we do not get

Table 1 The estimate $\sqrt{S_{zz}}$ (thousands of pounds per acre) of the standard deviation of z, by model and sampling method; m = number of clusters, p = percent of trees sampled per cluster.

Sampling Method		Model									
m	p	6	13	15	17	20	22	25	27	30	32
10	10	9.72	6.63	10.31	6.74	9.63	6.58	11.82	7.70	11.05	7.54
	20	8.41	5.96	8.48	6.07	8.52	5.62	9.14	6.06	8.57	6.04
	30	7.96	5.58	7.85	5.71	7.98	5.32	8.64	6.12	8.40	6.13
	40	7.71	5.46	7.51	5.54	7.63	5.27	8.09	5.90	8.23	6.01
	50	7.71	5.30	7.42	5.41	7.52	5.14	8.11	5.90	8.24	6.02
	60	7.54	5.16	7.36	5.22	7.40	5.01	7.85	5.78	7.84	5.82
	70	7.52	5.13	7.33	5.18	7.42	5.04	8.04	5.63	8.04	5.65
	80	7.57	5.13	7.42	5.17	7.46	5.00	8.39	5.69	8.36	5.56
	90	7.51	4.99	7.44	5.05	7.39	4.90	8.55	5.57	8.45	5.41
	100	7.53	4.90	7.42	4.97	7.37	4.86	8.55	5.44	8.26	5.29
15	10	7.26	4.91	7.45	5.47	6.94	5.43	8.03	5.87	7.52	5.61
	20	6.55	4.76	6.73	4.97	6.56	4.94	7.34	5.32	6.77	5.28
	30	6.29	4.68	6.31	4.80	6.40	4.83	6.59	4.94	6.25	5.06
	40	6.18	4.71	6.22	4.83	6.16	4.82	6.79	4.98	6.51	4.95
	50	6.15	4.62	6.18	4.78	6.12	4.71	6.57	5.00	6.46	4.97
	60	6.02	4.47	5.99	4.62	6.00	4.60	6.41	4.93	6.32	4.86
	70	6.01	4.44	6.01	4.58	6.05	4.56	6.43	4.92	6.41	4.82
	80	6.00	4.45	6.01	4.57	6.01	4.55	6.35	4.88	6.54	4.82
	90	6.06	4.35	6.07	4.46	6.06	4.49	6.47	4.76	6.63	4.74
	100	6.09	4.33	6.05	4.49	6.05	4.54	6.47	4.78	6.64	4.68
20	10	5.62	4.09	5.59	4.32	5.59	4.38	5.79	4.67	5.65	4.30
	20	5.21	3.60	4.91	3.60	5.03	3.82	5.36	3.68	5.14	4.06
	30	4.95	3.62	5.06	3.57	5.03	3.69	5.47	3.64	5.19	3.77
	40	4.87	3.61	4.90	3.56	4.80	3.71	5.37	3.70	5.13	3.83
	50	4.76	3.50	4.75	3.46	4.66	3.56	5.19	3.71	5.00	3.78
	60	4.77	3.45	4.71	3.42	4.64	3.47	5.15	3.69	4.99	3.78
	70	4.74	3.36	4.71	3.36	4.62	3.41	5.01	3.63	5.00	3.69
	80	4.74	3.36	4.83	3.35	4.70	3.47	4.95	3.60	5.07	3.65
30	10	5.08	3.56	5.21	3.87	4.88	3.44	5.25	4.07	5.14	3.73
	20	4.58	3.16	4.49	3.21	4.37	3.24	4.60	3.37	4.55	3.53
	30	4.35	3.17	4.33	3.19	4.27	3.34	4.48	3.28	4.47	3.34
	40	4.21	3.17	4.05	3.18	4.02	3.30	4.07	3.24	4.09	3.38
	50	4.00	3.02	3.99	3.02	3.93	3.16	4.06	3.12	3.88	3.21
50	10	4.20	2.79	3.94	2.85	3.86	2.85	3.91	2.97	3.83	3.08
	20	3.68	2.56	3.49	2.52	3.53	2.47	3.62	2.60	3.43	2.86
	30	3.31	2.49	3.26	2.46	3.33	2.50	3.32	2.52	3.22	2.61

much more precise point estimates by using the slightly more complicated weighted procedures. However, $\sqrt{S_{zz}}$ is generally slightly higher for modified (MLS or MWLS) versions of a model than for unmodified (OLS or OWLS) versions of the same model. In other words, modified models are slightly less precise than analogous unmodified models. This difference is most noticeable with 10 cluster samples, and decreases as the number of clusters sampled increases. Although the difference in most cases is not statistically discernible (Gillespie 1985), it is nonetheless interesting that this trend is so consistent. One possible explanation for this trend is that the range of individual tree measurements y, d, and h is relatively larger than the range of average summed cluster variables u and v. This trend did not occur in earlier studies by Briggs

and Cunia (1982) and Kotimaki and Cunia (1981).

Analysis of $\sqrt{\bar{v}}$, the estimate of $\sqrt{\sigma_{zz}}$

The statistic \bar{v} is the estimate of σ_{zz} as calculated under the assumptions of each given model. We will focus on the standard deviation $\sqrt{\bar{v}}$, which is calculated as the square root of the arithmetic average of the 100 estimates V_{ij} of the variance of z_{ij} (one model applied to 100 samples). Table 2 shows $\sqrt{\bar{v}}$ for all 360 model-sampling method combinations in the study.

The estimate $\sqrt{\bar{v}}$ decreases as average sample size increases. For the unmodified regression models, $\sqrt{\bar{v}}$ seems directly dependent upon the number of trees in the sample size, regardless

Table 2 The estimate \sqrt{V} (thousands of pounds per acre) of the standard deviation of z , by model and sampling method. m = number of clusters, p = percent of trees sampled per cluster.

Sampling Method		Model									
m	p	6	13	15	17	20	22	25	27	30	32
10	10	9.66	6.46	7.35	5.05	8.29	6.19	10.31	6.13	12.51	7.80
	20	8.45	5.96	5.26	3.80	5.60	4.31	9.45	5.63	11.27	6.78
	30	7.97	5.76	4.50	3.21	4.51	3.46	8.85	5.43	9.88	6.21
	40	7.64	5.49	3.89	2.80	3.87	2.97	8.56	5.19	9.55	6.05
	50	7.54	5.52	3.45	2.52	3.44	2.66	8.13	5.33	9.36	6.04
	60	7.39	5.46	3.12	2.29	3.14	2.42	7.98	5.20	9.21	6.10
	70	7.38	5.38	2.90	2.13	2.89	2.23	7.94	5.30	9.02	6.06
	80	7.23	5.32	2.69	1.99	2.71	2.08	7.77	5.16	8.82	6.00
	90	7.18	5.29	2.55	1.89	2.55	1.96	7.95	5.13	8.90	5.96
	100	7.16	5.27	2.42	1.81	2.41	1.86	7.87	5.13	8.82	5.87
15	10	7.92	5.54	5.92	4.29	6.38	4.82	7.84	5.58	9.32	6.32
	20	6.76	4.92	4.33	3.17	4.34	3.36	7.21	5.01	7.95	5.43
	30	6.36	4.72	3.63	2.65	3.54	2.72	6.71	4.71	7.26	5.04
	40	6.10	4.55	3.12	2.30	3.05	2.35	6.37	4.48	6.95	4.85
	50	6.03	4.48	2.79	2.05	2.71	2.10	6.19	4.38	6.80	4.77
	60	5.93	4.44	2.53	1.87	2.48	1.92	6.05	4.29	6.71	4.74
	70	5.93	4.37	2.35	1.73	2.30	1.77	6.03	4.25	6.65	4.70
	80	5.83	4.34	2.18	1.63	2.15	1.66	5.86	4.20	6.53	4.68
	90	5.77	4.29	2.06	1.54	2.03	1.56	5.84	4.13	6.51	4.64
	100	5.75	4.28	1.97	1.48	1.91	1.48	5.81	4.13	6.44	4.58
20	10	6.85	4.92	5.16	3.77	5.49	4.15	6.68	4.86	7.71	5.45
	20	5.90	4.27	3.79	2.77	3.76	2.88	5.99	4.31	6.58	4.67
	30	5.59	4.07	3.17	2.31	3.08	2.35	5.73	4.09	6.20	4.39
	40	5.38	3.91	2.73	2.01	2.65	2.04	5.52	3.91	5.95	4.23
	50	5.30	3.86	2.44	1.79	2.36	1.82	5.35	3.85	5.81	4.16
	60	5.20	3.81	2.21	1.63	2.16	1.67	5.22	3.77	5.72	4.11
	70	5.17	3.77	2.04	1.51	2.01	1.54	5.16	3.72	5.67	4.07
	80	5.08	3.74	1.92	1.43	1.87	1.45	5.00	3.67	5.57	4.05
30	10	5.63	4.04	4.29	3.18	4.32	3.36	5.37	4.01	5.89	4.39
	20	4.84	3.48	3.11	2.32	3.02	2.34	4.70	3.50	5.07	3.74
	30	4.54	3.27	2.58	1.90	2.47	1.92	4.41	3.27	4.77	3.52
	40	4.38	3.16	2.23	1.66	2.14	1.66	4.27	3.15	4.63	3.41
	50	4.30	3.10	1.99	1.48	1.91	1.49	4.15	3.09	4.55	3.34
50	10	4.41	3.20	3.44	2.58	3.33	2.61	4.32	3.25	4.52	3.41
	20	3.80	2.73	2.45	1.82	2.35	1.83	3.70	2.77	3.92	2.93
	30	3.57	2.56	2.04	1.49	1.91	1.50	3.46	2.57	3.69	2.76

of the sampling method. For the ratio estimators and modified regression models, all of which account for the cluster effect of sampling, \sqrt{V} behaves much like $\sqrt{S_{zz}}$; specifically, \sqrt{V} decreases with an increase in the number of clusters sampled, as well as with increased value of p (up to 30 to 40 percent). For these models, the effect of increasing the number of clusters does more to reduce \sqrt{V} than increasing percent subsampling. Estimates of \sqrt{V} from models which use only d are generally 1.3 to 1.5 times as large as estimates from models which use d and h .

With small sample sizes, \sqrt{V} as estimated by weighted least squares models is generally slightly higher (10 to 30 percent) than \sqrt{V} as

estimated by unweighted least squares regression models for similar sampling methods. This difference gets smaller as the number of clusters and/or the percent subsampling increases. The estimates of \sqrt{V} from modified regression models are 30 to 50 percent higher than estimates from unmodified regression models. The estimates of \sqrt{V} from ratio estimator models are larger than those from unmodified regressions, but smaller than those from modified regressions (which use the same independent variable(s)).

Table 3 The ratio $\sqrt{\bar{V}/S_{ZZ}}$ of the two different estimates of the standard deviation of z, by model and sampling method. m = number of clusters sampled, p = percent of trees sampled per cluster.

Sampling Method		Model									
m	p	6	13	15	17	20	22	25	27	30	32
10	10	0.99	0.97	0.71	0.75	0.85	0.94	0.87	0.80	1.13	1.04
	20	1.00	1.00	0.62	0.63	0.66	0.77	1.03	0.93	1.32	1.12
	30	1.00	1.03	0.57	0.56	0.57	0.65	1.02	0.89	1.18	1.01
	40	0.99	1.01	0.52	0.51	0.51	0.56	1.06	0.88	1.16	1.01
	50	0.98	1.04	0.46	0.47	0.46	0.52	1.00	0.90	1.14	1.00
	60	0.98	1.06	0.42	0.44	0.42	0.48	1.02	0.90	1.18	1.05
	70	0.98	1.05	0.40	0.41	0.39	0.44	0.99	0.94	1.12	1.07
	80	0.96	1.04	0.36	0.38	0.36	0.42	0.93	0.91	1.06	1.08
	90	0.96	1.06	0.34	0.37	0.35	0.40	0.93	0.92	1.05	1.10
	100	0.95	1.07	0.33	0.36	0.33	0.38	0.92	0.94	1.07	1.11
15	10	1.09	1.13	0.79	0.78	0.92	0.89	0.98	0.95	1.24	1.13
	20	1.03	1.03	0.64	0.64	0.66	0.68	0.98	0.94	1.17	1.03
	30	1.01	1.01	0.57	0.55	0.55	0.56	1.02	0.95	1.16	1.00
	40	0.99	0.96	0.50	0.48	0.50	0.49	0.94	0.90	1.07	0.98
	50	0.98	0.97	0.45	0.43	0.44	0.45	0.94	0.88	1.05	0.96
	60	0.99	0.99	0.42	0.40	0.41	0.42	0.94	0.87	1.06	0.97
	70	0.99	0.98	0.39	0.38	0.38	0.39	0.94	0.86	1.04	0.97
	80	0.97	0.98	0.36	0.36	0.36	0.36	0.92	0.86	1.00	0.97
	90	0.95	0.99	0.34	0.35	0.33	0.35	0.90	0.87	0.98	0.98
	100	0.94	0.99	0.33	0.33	0.32	0.33	0.90	0.86	0.97	0.98
20	10	1.22	1.20	0.92	0.87	0.98	0.95	1.15	1.04	1.37	1.27
	20	1.13	1.19	0.77	0.77	0.75	0.75	1.12	1.17	1.28	1.15
	30	1.13	1.12	0.63	0.65	0.61	0.64	1.05	1.12	1.19	1.16
	40	1.10	1.08	0.56	0.56	0.55	0.55	1.03	1.06	1.16	1.11
	50	1.11	1.10	0.51	0.52	0.51	0.51	1.03	1.04	1.16	1.10
	60	1.09	1.10	0.47	0.48	0.47	0.48	1.01	1.02	1.15	1.09
	70	1.09	1.12	0.43	0.45	0.43	0.45	1.03	1.02	1.13	1.10
	80	1.07	1.11	0.40	0.43	0.40	0.42	1.01	1.02	1.10	1.11
30	10	1.11	1.13	0.82	0.82	0.89	0.98	1.02	0.98	1.15	1.18
	20	1.06	1.10	0.69	0.72	0.69	0.72	1.02	1.04	1.12	1.06
	30	1.04	1.03	0.59	0.60	0.58	0.57	0.99	1.00	1.07	1.05
	40	1.04	1.00	0.55	0.52	0.53	0.50	1.05	0.97	1.13	1.01
	50	1.08	1.03	0.50	0.49	0.49	0.47	1.02	0.99	1.17	1.04
50	10	1.05	1.15	0.87	0.91	0.86	0.92	1.10	1.10	1.18	1.11
	20	1.03	1.07	0.70	0.72	0.66	0.74	1.02	1.06	1.14	1.02
	30	1.08	1.03	0.62	0.60	0.57	0.60	1.04	1.02	1.14	1.06

Analysis of the ratio $\sqrt{\bar{V}/S_{ZZ}}$

The ratio $\sqrt{\bar{V}/S_{ZZ}}$ illustrates the difference between the assumed and the actual precision of the various models. Table 3 lists this ratio for all 360 model-sampling method combinations.

For the ratio of mean estimators and modified regression, the ratio $\sqrt{\bar{V}/S_{ZZ}}$ is approximately 1.0 or slightly higher, staying constant regardless of changes in the sampling method or sample size. For unmodified regression models, $\sqrt{\bar{V}/S_{ZZ}}$ is generally between .4 and 1.0, decreasing with increases in percent subsampling from a fixed number of clusters, and increasing with an

increase in the number of clusters sampled for a fixed percentage of trees. The net effect is a slight decrease in the ratio as the average sample size increases, indicating that \bar{V} decreases relative to S_{ZZ} .

In general, the ratio $\sqrt{\bar{V}/S_{ZZ}}$ is larger for weighted regression models than for similar but unweighted regression models. The ratio is also much larger for modified regression models (approximately 1.0 or higher) than for similar but unmodified regression models (from .4 to 1.0).

In all cases of unmodified regression models, the ratio is less than 1.0. If S_{ZZ} is

accepted as a better (and unbiased) estimate of σ_{zz} , then V underestimates σ_{zz} in unmodified models. This is expected, since the assumptions of independence among sample elements is violated.

These relationships are emphasized by the analysis of confidence interval reliability. Unmodified regression models yield 95 percent confidence intervals which were only correct 40 percent of the time. Ratio estimators and modified unweighted regression models yield 95 percent confidence intervals which are correct approximately 95 percent of the time (except when samples came from 10 clusters). Modified weighted regression models yield 95 percent confidence intervals which are 95 to 100 percent correct, implying that they may be too large and that the weighted modified procedure overestimates the error of the estimate.

It is important to note that the study population was simulated by a weighted, unmodified regression model with conditional variance set proportional to $d^4 h^2$. In the modified weighted regression models, the cluster biomass was defined as $u = \sum y$, with the variance of u assumed to be proportional to the sum of the a_i^2 (with $a_i^2 = d^4$ or $d^4 h^2$). A better assumption (based on the success of the modified unweighted regression models) seems to be that the conditional variance of u is proportional to the number of sample trees in the cluster, n_h , even when the individual tree biomass conditional variance is known to be proportional to $d^4 h^2$ (Gillespie, 1985).

Summary Comments

The main objectives of this study were to verify the results of Briggs and Cunia (1982) and Kotimaki and Cunia (1981) concerning the estimation of biomass tables by cluster sampling. These earlier studies found that the point estimates of biomass were approximately equally accurate whether or not the estimation procedure considered the cluster effect of sampling. However, they felt (but could not prove) that ignoring the cluster effect caused a large underestimation of the error of such point estimates.

Our conclusions generally agree with these results. The ratio-of-means estimator and parabolic least squares regression models were practically unbiased estimators of μ , especially when samples were selected from at least 15 clusters. There was no statistically discernible difference between modified and unmodified procedures in calculating point estimates of biomass.

The estimate S_{zz} of error of the point estimates was (i) much lower for models using both d and h than for models using d alone; (ii) roughly equal for weighted and unweighted versions of a similar regression model; and (iii) slightly higher for modified (for cluster effect) versions of regression models than for unmodified versions of the same model.

For all models, S_{zz} decreased with both an increase in the number of clusters sampled for a fixed average sample size, and with increased percentage p subsampling from a fixed number of clusters. This decrease tapers off after 30 to 40 percent of subsampling.

The estimate V of the error of the point estimates reflects the model assumptions concerning variance estimation. For unmodified regression models, V was consistently less than S_{zz} , which implies that ignoring the cluster effect will lead to underestimation of the error, or overestimation of precision. For modified models and the ratio of means estimators which considered the cluster effect, V was approximately equal to or slightly greater than S_{zz} , except for modified weighted regression models when V was consistently larger (by up to 25 percent) than S_{zz} . This overestimation of the true error by modified weighted regression models was not reported in the earlier studies. The consistency of this behavior over the variety of sampling methods studied implies that the assumptions of the modified weighted regression estimators concerning variance estimation may not be correct for our study population. This was emphasized by the behavior of confidence interval reliability, where counts of confidence intervals which included the known population mean μ (out of sets of 100 trials) were consistently greater than expected. Unweighted modified regression models yielded the best estimates of error, implying that the conditional variance of the cluster sum of biomass is proportional to the number of trees in the cluster, rather than the sum of d^4 or $d^4 h^2$.

The present study was limited to simple random samples of clusters and of a fixed proportion of trees within clusters. Future studies may have to consider (i) selecting trees from clusters with probability proportional to some measure of tree size; (ii) consideration of other estimation procedures, such as generalized least squares or a hybrid regression which combines OLS point estimates and MLS error estimates; and (iii) estimation techniques applied to portions of samples, for example trees above a specified DBH.

Acknowledgements

This paper is based on the Master's thesis of the senior author. The research was funded by the Research Foundation of the State of New York, the United States Department of Agriculture-Forest Service, and the Department of Energy, grant number 23-524. We thank Ms. Sueh-Fang Hsu for permitting use of her simulated samples and allowing modification of one of her computer programs.

Literature Cited

Neter, J., W.; Wasserman, W.; Kutner, M. H. Applied linear regression models. Homewood, IL: Richard D. Irwin, Inc; 1983. 547 p.

Briggs, E. F.; Cunia, T. Effect of cluster sampling in biomass table construction: linear regression models. Canadian Journal of Forest Research 12: 255-263; 1982.

Cochran, W. G. Sampling techniques. 3d ed. New York: John Wiley & Sons; 1977. 428 p.

Cunia, T. On tree biomass tables and regression: some statistical comments. In: 1979 Forest resource inventories workshop proceedings, W. E. Frayer, (Ed.) Colorado State University, Fort Collins, CO; 1979: v. II, 629-642.

Cunia, T.; Gillespie, A. J. Cluster sampling and construction of biomass tables: Results of a simulation study. In: Proceedings, third annual southern forest biomass workshop, 1985 March 12-14; University of Florida, Gainesville, FL.

Cunia, T.; Michelakackis, J. A method to construct a forest biomass population model. In: Proceedings, Resources inventories for monitoring changes and trends; 1983; J. F. Bell and T. Atterbury, (Eds.); Oregon State University Corvallis, OR; 1983: 558-562.

Cunia, T.; Michelakackis, J. A Monte Carlo technique for generating total heights of forest trees. Syracuse, NY: SUNY College of Environmental Science and Forestry; 1984a; School of Forestry Miscellaneous Publication Number 4 (ESF 84-018). 24 p.

Cunia, T.; Michelakackis, J. Constructing forest biomass populations for simulated sampling. Syracuse, NY: SUNY College of Environmental Science and Forestry; 1984b; School of Forestry Miscellaneous Publication Number 5 (ESF 84-019). 46 p.

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

Draper, N. R.; Smith, H. Applied regression analysis. 2d ed. New York: John Wiley and Sons; 1981.

Gillespie, Andrew J.R. Estimation of biomass tables by cluster sampling: a simulation study. M.S. thesis, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1985. 302 p.

Kotimaki, T. A.; Cunia, T. Effect of cluster sampling in biomass table construction: ratio estimators models. Canadian Journal of Forest Research 11: 475-486; 1981.

ERROR OF BIOMASS REGRESSIONS: SAMPLE TREES

SELECTED BY STRATIFIED SAMPLING

Alexandros Arabatzis and Tiberius Cunia

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

It is common to select sample trees for biomass tables construction by stratified, two-stage cluster sampling whereby (i) several clusters (plots) of trees are selected at random from each of a given number of strata and (ii) from each sample cluster a certain percent of trees is selected at random to be measured for biomass, diameter and height. The common procedure to estimate a biomass regression function applicable to the entire forest area (containing all strata) is to use the least or weighted least squares method applied to the individual tree data irrespective of the stratum from which each tree has been selected. Simulation techniques were used to evaluate the validity of the inferences made when (i) the above procedure is used to estimate the biomass regression function and the average biomass per unit area of a given forest, or (ii) this procedure is replaced by alternative procedures that take into account the effect of the stratum from which the individual trees are being selected. It is shown that (i) the common estimation procedure is generally poor as it may lead to biased results and (ii) the proper estimation procedure is to calculate biomass regression functions separately by stratum.

Introduction

Cunia (1986) describes a computer simulation process whereby (i) samples of trees are selected by a variety of sampling techniques from a large population of forest trees, (ii) for each individual sample so selected, a variety of statistical procedures are used to estimate the regression function of tree biomass on diameter or diameter and height, (iii) each biomass regression of each individual sample is applied to the tree population values of diameter and height to estimate the mean biomass per acre and (iv) by comparing the estimates with the known value of the population mean biomass per acre, inferences are made about the probability behavior of the estimates; inferences about the effect, if any, of the sampling method and estimation procedure on the bias of the estimates, their precision and their estimated precision.

He has described three major sampling techniques. The first is the two-stage random sampling where m sample plots are selected from the population of 667 non-empty plots (clusters of trees) by simple random sampling without replace-

ment and the trees from the selected plots are subsampled by a variety of sampling procedures. A second sampling method, defined as two-phase, two-stage random sampling, generates samples consisting of two, two-stage random samples; a first phase, relatively small sample where the trees are measured for diameter, height and biomass and a second phase, relatively large sample with the trees measured for diameter and height but not biomass. The third sampling method, denoted here as stratified, two-stage sampling is similar to the first; the difference is that the m sample plots of the first stage are selected by stratified, not simple random sampling.

Cunia and Gillespie (1985), Gillespie (1985), Gillespie and Cunia (1986) and Michelakackis and Cunia (1985, 1986) discuss in more detail the simulation processes and the results obtained by the first two sampling methods. In the present paper we shall discuss the results obtained by the third method. The detailed results are reported by Arabatzis (1986). More specifically, we shall describe the simulation process applied to stratified two-stage sampling and its five components, namely, the population being simulated, the sampling method, the estimation procedures for the biomass regression function, the application of the biomass regression to estimate the mean biomass per acre and finally, the analysis of the results obtained. Not to repeat processes already described or conclusions already reached, we shall refer heavily to the above-mentioned papers.

Forest Tree Populations Being Simulated

The basic tree population, hereby called Population 1 has been constructed by a procedure described in a series of papers by Cunia and Michelakackis (1983, 1984a,b) and Cunia, Michelakackis and Lee (1984) and summarized in a paper by Cunia (1986). For more details the interested reader is referred to these papers. It suffices to state here that (i) the basic data used to construct the tree population consists of actual field measurements of diameter at breast height (d), merchantable height (h_m) and species, among other things, performed on 22753 trees of merchantable size ($d > 5$ inches) contained in 927 one-fifth acre plots (260 of which are empty, that is, without merchantable size trees) selected from New York State forest lands and (ii) the total height and biomass of each individual tree was generated by Monte Carlo techniques that took into account the effect of diameter, merchantable height, species, site quality, geographical region, individual plot and finally, the random variation of tree height or biomass about its own (conditional) expected value.

To apply stratified sampling, one must use strata that are simple to define and sample, and sufficiently different from each other to make the sampling efficient. For simplicity, the sample plots were classified into three strata by the geographical region in which they happened to fall. Unfortunately, this resulted in strata

that were very similar to each other. Preliminary samples with plots selected by an allocation that was far from proportional, generated unstratified means (which ignored the plot stratification) that were not significantly different from the true mean (biomass per acre) of the population. This stratification obviously was not appropriate for the objectives of our study. Consequently, changes were required in the population tree biomass to make the differences between strata sufficiently large. This was accomplished by increasing (or decreasing) the biomass of each individual tree from a given stratum by an amount calculated by a formula of the form $(a+bd^2)$, where the constants a and b were stratum specific and d was the tree diameter.

More specifically, we have defined (i) the new Population 2 by using the values (a,b) equal to (238, -2.46), (60, .60) and (-11, -1.34) for the trees of stratum 1, 2, and 3 respectively, (ii) the new Population 3 by using (200, .40) for stratum 1 and (-11, -1.34) for stratum 3, with the biomass of the trees of stratum 2 remaining unchanged and (iii) the new Population 4 where the values (a,b) were those equal to (150, .30), (10, -.50) and (200, 1.50) for stratum 1, 2 and 3 respectively. As it will be later seen, the three strata of the new three populations become sufficiently different.

Stratified Two-Stage Sampling Procedure

The first stage of the sampling procedure consists of m_1 , m_2 and m_3 non-empty sample plots (clusters of trees) selected by simple random sampling from the 233, 188 and 246 non-empty plots of stratum 1, 2, and 3 respectively. There are also 64, 119 and 77 empty plots (without merchantable size trees) in these strata, but there was no point in selecting such plots in the sample. They were considered, however, when the estimate of the mean biomass per acre was calculated. In the second stage, a fixed percentage p of trees from each plot of the first stage sample was selected by simple random sampling without replacement. Because the multiplication of the percentage p with the number n_i of trees in the i-th plot is not necessarily an integer, a random number R from .00 to .99 was used to decide whether an additional sample tree is to be selected from the plot (when $R <$ fractional part of the product $n_i p$) or no additional tree is to be selected otherwise.

For the purpose of the present study forty sets of 100 independent samples were selected by computer from each of the three strata. These forty sets represent forty different combinations of (i) first stage number m of sample plots ($m = 2, 5, 10, 15, 20, 30$) and (ii) second stage percentage p of sample trees ($p = 5, 10, 15, 20, 40, 60, 100$). Not all combinations of m and p values above were used, and the resulting sample size of the combinations used varied from 7 to 1020 trees. The specific combinations of m and p that were used are listed in Arabatzis (1986); they are not given here.

A stratified sample is now defined as a combination of three samples, one from each stratum. As there are 4000 samples from each stratum (40 sets of 100 simulated samples) there are theoretically a total of $4000^3 = 64,000,000,000$ possible stratified samples. To reduce the number of these samples to a manageable size, we have decided to consider stratified samples that (i) use the same percentage p in all strata, (ii) are generated by the same simulation run (that is, they have the same simulation run number) and (iii) consist of arbitrarily selected values m_1 , m_2 and m_3 . The specific types of stratified samples used will be given later.

Estimation Procedures for the Biomass Regressions

The various procedures to estimate the tree biomass regression functions are defined as combinations of (i) assumed form of the regression function, (ii) type of least squares method used and (iii) way by which the stratification of sample tree data is taken into account.

Two basic forms of regression function were used,

$$\text{form 1: } \hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2, \text{ and}$$

$$\text{form 2: } \hat{y} = \beta_1 + \beta_2 d^2 + \beta_3 d^2 h$$

where y = tree biomass and, as defined before, d = tree diameter and h = height. The selection of these forms was based on (i) the conclusions reached by Cunia and Gillespie (1985), Gillespie and Cunia (1986) and Michelakackis and Cunia (1985, 1986) and (ii) the fact that a term of the form bd^2 was added to each tree biomass of Populations 2, 3 and 4.

Four least squares methods were used, the ordinary least squares (OLS), ordinary weighted least squares (OWLS), modified least squares (MLS) and modified weighted least squares (MWLS). These methods are described in more detail in Cunia and Gillespie (1985), Gillespie and Cunia (1986), Michelakackis and Cunia (1985, 1986), Cunia (1986) and Arabatzis (1986). This detailed description is not repeated here. It suffices to say that (i) OLS is the usual least squares regression procedure applied to individual tree data without reference to the plot the tree was selected from, (ii) OWLS is the usual weighted least squares method where the conditional variance of y given d is assumed to be proportional to d^4 and the conditional variance of y given d and h is assumed to be proportional to $d^4 h^2$ and (iii) MLS and MWLS are the usual least and weighted least squares methods modified so as to take the cluster (plot) effect into account; these modifications consisting of the application of the weighted least squares method to the plot (not tree) data with the assumptions about the weights to use varying from MLS to MWLS.

We have defined nine procedures to take (or not to take) into account the effect of stratification, when the data of the sample trees are being used. These procedures are described in

detail by Arabatzis (1986) and for more information, the reader is referred to his work. However, because they are essential to the interpretation of the results obtained and conclusions reached from the analysis of these results, the nine procedures are summarized below.

Procedure 1 - The trees are assumed to have been selected by unstratified sampling and, thus, one common regression function is estimated, and then applied, to all three strata. This is the usual procedure used by foresters when one regression function is desired for all strata and stratified sampling is used to select the sample trees in order to insure a better representation of the trees from each stratum.

Procedure 2 - Stratified sampling is assumed, with the trees from one stratum selected independently from those of another stratum. Three biomass regression functions, one for each stratum, are estimated independently of each other and each regression is applied to its own stratum when the estimate of the mean biomass per acre is calculated. This is the right procedure to use when the sample trees were selected by the classical method of stratified random sampling.

Procedure 3 - Same as Procedure 2 but the three biomass regressions are calculated simultaneously by weighted least squares regression techniques that use dummy variables to (i) test null hypotheses about similarities between the regression functions of various strata and (ii) calculate regression coefficients that are common to several regressions, when these regression coefficients are not significantly different from each other. The least squares with dummy variable techniques are fully described by Cunia (1973) and the interested reader should refer to him for more details.

Procedure 4 - The three stratum biomass regression functions of Procedure 2 are first calculated as usual, and then, a single biomass regression function, applicable to all strata is defined as the weighted regression

$$\hat{Y} = w_1 r_1 + w_2 r_2 + w_3 r_3$$

where r_i is the regression of stratum i and the weight w_i is defined as the ratio of the total area of stratum i to the total area of all strata, for $i = 1, 2, 3$.

Procedure 5 - Similar to Procedure 4 but the weights w_i are defined as the ratio of the total number of trees in stratum i to the total number of trees in all strata, $i = 1, 2, 3$.

Procedure 6 - Similar to Procedure 4 but the three biomass regression functions that were averaged were calculated by Procedure 3 and not Procedure 2.

Procedure 7 - Similar to Procedure 6 but the weights w_i are those of Procedure 5 (based on number of trees) and not those of Procedure 4 (based on stratum areas).

Procedure 8 - A single biomass regression function is calculated for all three strata by the weighted least squares method, where the weights used are defined as functions of (i) area of stratum, (ii) sample size within a stratum and (iii) the usual tree values of d or d and h . More specifically, the weighted least squares method was applied as if, for each individual tree, the conditional variance of the biomass is (i) proportional to stratum area, (ii) inversely proportional to sample size within stratum and (iii) proportional to d^4 or $d^4 h^2$ as the case may be.

Procedure 9 - Similar to Procedure 8, but the weight of each individual tree is made proportional to stratum number of trees rather than stratum area.

Combining the form of the regression function (2) with the least squares method (4) and the procedure to take the tree stratification into account (9) yields a total of 72 procedures for the estimation of the biomass regression functions. This means that, for each simulated sample, we have calculated a total of 72 biomass regressions. To simplify the discussion we shall identify the three criteria of the estimation procedure as (i) form of regression function (ii) least squares estimation approach and (iii) procedure number (to take the tree stratification into account).

Note that, from among the nine procedures we have described, the second procedure is one which is theoretically unbiased, provided of course that the estimators of the means μ_i within stratum i are themselves unbiased. Consequently procedure 2 may be taken as the one to be compared with all of the remaining eight procedures.

Estimation Procedure for the Mean Biomass per Acre

To calculate z , the estimate of the overall mean biomass per acre of the entire population, we shall use the stratified sampling formula

$$z = Q_1 z_1 + Q_2 z_2 + Q_3 z_3$$

where, for stratum $i = 1, 2, 3$,

A_i = area of stratum i (acres)
 $A = A_1 + A_2 + A_3$ = total area (in acres)
 = total area (in acres) of the 927 one-fifth acre plots of our population
 $Q_i = A_i/A$ = relative size of stratum i , and
 z_i = estimate of the mean biomass per acre μ_i of stratum i .

Let us write now the biomass regression function of stratum i as

$$\hat{y} = b_{i1} x_1 + b_{i2} x_2 + b_{i3} x_3$$

where

$$x_1 = 1 \text{ for all trees}$$

$x_2 = d$ or d^2 for regression function form 1 or 2 respectively, and

$x_3 = d^2$ or d^2h for regression function form 1 or 2 respectively.

If the totals of the variables y , x_1 , x_2 and x_3 for stratum i are written as

τ_i = total of tree biomass y in stratum i

τ_{i1} = total of values x_1 in stratum i

= total number of trees in stratum i

τ_{i2} = total of values x_2 in stratum i

= sum of tree diameters d or squared tree diameters d^2 for regression function form 1 or 2 respectively, and

τ_{i3} = total of values x_3 in stratum i

= sum of squared tree diameters d^2 or crossproducts d^2h for regression function form 1 or 2 respectively,

then, the corresponding averages on a per acre basis are

τ_i/A_i = mean biomass per acre in stratum i

τ_{i1}/A_i = mean number of trees per acre in stratum i

τ_{i2}/A_i = mean sum of d or d^2 per acre as the case may be and

τ_{i3}/A_i = mean sum of d^2 or d^2h per acre as the case may be

Using the above notation it is easy to show that

(1) the estimate z_i of μ_i can be written as

$$z_i = b_{i1}(\tau_{i1}/A_i) + b_{i2}(\tau_{i2}/A_i) + b_{i3}(\tau_{i3}/A_i)$$

and

(2) the estimate z of the overall mean μ is

$$\begin{aligned} z &= Q_1 z_1 + Q_2 z_2 + Q_3 z_3 \\ &= (1/A)(b_{11}\tau_{11} + b_{12}\tau_{12} + b_{13}\tau_{13} + b_{21}\tau_{21} \\ &\quad + \dots + b_{33}\tau_{33}) \\ &= [B]'[\mu_x] \end{aligned}$$

where $[]$ and $[]'$ denotes matrices and transpose matrices

$$[B]' = [b_{11} \ b_{12} \ b_{13} \ b_{21} \ b_{22} \ b_{23} \ b_{31} \ b_{32} \ b_{33}]$$

and

$$[\mu_x]' = [(\tau_{11}/A) \ (\tau_{12}/A) \ \dots \ (\tau_{33}/A)]$$

Because $[\mu_x]$ is known without error, and if the covariance matrix of $[B]$ is estimated by $[S_{BB}]$, then, the variance of z is estimated by the expression

$$V = [\mu_x]' [S_{BB}] [\mu_x]$$

If the basic assumptions of the sampling and estimation procedures are strictly satisfied, z and V are unbiased estimators. As these assumptions are never satisfied, both z and V are biased. The objectives of this study is to find, among other things, to what extent z and V are biased.

Analysis Procedure

Each simulated sample may be viewed as a random experiment. For a given sampling procedure, sample size and allocation, the 100 simulated samples may also be viewed as 100 random outcomes, the result of 100, identically defined and statistically independent random experiments. Each random experiment generates 72 sets of random variables, one set for each estimation procedure. Each set contains (i) two random variables z and V (as defined in the previous section) that are taken as estimators of the mean biomass per acre μ and the variance of z (as an estimator of μ) respectively, (ii) two random intervals ($z \pm t\sqrt{V}$), with $t = 2.0$ and $t = 2.6$ and (iii) two multinomial random variables that take the values -1 , 0 and 1 , depending on whether the μ falls below, within or above the two random intervals above.

Because, for each estimation procedure and sampling method, sample size and allocation we have 100, identically distributed and statistically independent random variables z , we can use their average \bar{z} and variance S_{zz} as estimators of the mean μ_z and variance σ_{zz} . As z is taken as an estimator of μ , (i) the sample statistic $(\bar{z} - \mu)$ may be used as an estimator of the bias of z , (ii) the sample statistic $S_{zz} = S_{zz}/100$ is an estimator of the variance of the bias $(\bar{z} - \mu)$ of z (as well as estimator of the variance of \bar{z}) and (iii) $t = (\bar{z} - \mu)/\sqrt{S_{zz}}$ can be used as a statistic to test the null hypothesis that the bias of z is equal to zero.

The variance of z is also estimated by the 100 sample values V as well as their average \bar{V} . This last value would normally be an unbiased estimator of the true variance of z whenever all the assumptions of the statistical model used in the corresponding sampling and estimation procedure are strictly satisfied. Another estimator of the same true variance of z , this time unbiased, is the statistic S_{zz} defined above. Consequently, the ratio V/S_{zz} (or its square root), may be taken as an estimator of the relative amount by which V underestimates or overestimates the variance of z . Another measure of the goodness of V as an estimator of the variance of z is provided by the number of times (out of the 100 times) that the value μ fell below, within or above the 95 and 99 percent confidence intervals.

All of these values for all stratified sample sizes and estimation procedures considered in this study are listed in a series of tables by Arabatzis (1986). As an example, we have shown in Tables 1, 2, 3, and 4, for populations 1, 2, 3 and 4 respectively and for the 72 estimation procedures, the statistics \bar{z} , $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, \sqrt{V} , $\sqrt{V/S_{zz}}$, t and the number of times μ fell below, within and above the 95 and 99 percent confidence intervals, for the stratified sampling method with $m_1=30$, $m_2=20$, $m_3=10$ and $p=15$ percent.

The procedure to analyze the tables is to a large extent subjective. The conclusions were drawn by an ocular analysis of the values of \bar{z} , $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, $\sqrt{V/S_{zz}}$, etc. as listed in the table or as plotted in a set of graphs.

Analysis of Results - Part 1

To get an overall view of the type of results obtained, we have analyzed first, in relative detail, one specific stratified sample size and allocation, the case of $m_1=30$, $m_2=20$ and $m_3=10$ clusters selected at random from stratum 1, 2 and 3 respectively with 15 percent of their trees subsampled at random. The statistics \bar{z} , $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, \sqrt{V} , $\sqrt{V/S_{zz}}$, t and the number of times the confidence intervals fell above, over and below μ are listed by estimation procedure in Tables 1, 2, 3 and 4 for Populations 1, 2, 3 and 4 respectively. We shall successively analyze the bias of z , the precision and estimated precision of z and finally, the confidence intervals (that include the combined effect of the bias and estimated precision).

Analysis of the Bias of z

An ocular analysis of the bias $(\bar{z}-\mu)$ of Population 1 shows that, with the exception of the ordinary weighted least squares estimates by the regression form 2, the bias is not significantly different from zero; the values t are small, negative or positive values. And even when the bias is significant, the significance level is ordinarily above one percent. This is to be expected since the difference between the various strata is very small and any estimation procedure that gives too much weight to any given stratum could not unduly affect the bias. This seems to imply that, whenever the strata are sufficiently similar, the bias of the estimator z of μ would normally be close to, if not equal to zero.

The conclusions change, however, with the other three populations, where the strata were made, on purpose, to be different from each other. Starting with Procedure 1 (of a single unstratified biomass regression) the reader can verify that the sample bias is high in value and highly significant. Depending on the population considered, the bias is about 3, 7 and -5 percent of μ and the value of the test statistic t is about 10, 20 and -17. Consequently, the evidence strongly suggests that treating a stratified as an unstratified sample (when calculating the

biomass regression) is a poor procedure as it may seriously bias the estimates. This is to be expected on purely intuitive grounds.

With the Procedure 2 estimates, the results are different; the bias is small and not significantly different from zero. The one exception is with the estimates by the ordinary weighted least squares applied to the regression form 2. This is hard to explain. It may be something inherent to the way by which the estimates are calculated by the weighted least squares. It may be due to some specific (and unidentified) characteristic of our population. We do not know. But because (i) the weighted least squares estimates of the regression form 1 are statistically unbiased (ii) we expected statistically significant biased results to occur once in a while even where the procedure generating them is unbiased, and (iii) the value of the t -statistic is relatively low, about -2.34, we conclude that, in general Procedure 2 yields, as intuitively expected, unbiased results. The procedure estimates the mean μ_i of each stratum i separately and independently of the estimation of the mean μ_j of any other stratum j and then combines the three estimates z_1 , z_2 and z_3 into a single estimate of the overall mean μ by formulae which, from a theoretical point of view at least are unbiased.

Procedure 3 forces the regressions of the individual strata (of Procedure 2) to be parallel or identical for some simulated samples, when it is indeed known (in our case) that these regressions are not parallel or identical. This is a possible source of bias. As shown by an ocular analysis of Tables 2, 3 and 4, the results are mixed; the bias is not significantly different from zero for Population 2 but the bias is sometimes significant and sometimes not significant for the other two populations. This seems to imply that the third procedure may introduce a bias in the estimates, even though this bias is, most of the time relatively small, of the order of less than one percent of μ .

Procedures 4, 6 and 8, where a single regression function (common to all strata) is calculated by procedures using weights proportional to stratum areas (as measures of stratum sizes) yield high bias values that are significantly different from zero. Sometimes, the bias value may happen to be small and not significant. But this seems to occur at random and one cannot find a specific form of regression function or a least squares approach for which the bias would be consistently small. Things are different, however, when the measure of the stratum size is based on the total number of trees it contains. Then, the results are much better. Procedures 5 and 9 produce consistently small bias values which only occasionally are significantly different from zero, while for Procedure 7, the bias is significantly different from zero in 7 out of 24 cases. This last result is not unexpected since the single biomass regression used by Procedure 7 is based on three separate regressions which applied separately to the corresponding strata yield, in several instances, biased results. All this seems to imply that (i) using

TABLE 1: The statistics \bar{z} , $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, \sqrt{V} , the ratio \sqrt{V}/S_{zz} , t and the number of time μ fell below, within or above the 95% and 99% confidence intervals, as calculated by procedures 1, 2, 3, 4, 5, 6, 7, 8 and 9 for population 1; Sampling method: 15 percent of trees selected from 30, 20 and 10 plots of stratum 1, 2 and 3 respectively.

estimation model/ procedure	\bar{z}	$(\bar{z}-\mu)$	$\sqrt{S_{zz}}$	\sqrt{V}	\sqrt{V}/S_{zz}	t	number of times						
							95%			99%			
							b	w	a	b	w	a	
1	111	114.989	-0.56	3.569	2.551	0.71	-1.57	5	84	11	2	94	4
2	121	115.098	-0.45	3.433	2.508	0.73	-1.32	7	85	8	1	95	4
3	131	115.057	-0.49	3.720	3.608	0.97	-1.32	2	94	4	1	96	3
4	141	115.065	-0.48	3.573	3.887	1.09	-1.36	0	97	3	0	98	2
5	211	115.770	0.22	2.575	1.887	0.73	0.86	11	84	5	4	93	3
6	221	114.933	-0.61	2.464	1.964	0.80	-2.50	2	87	11	2	94	4
7	231	115.745	0.19	2.711	2.663	0.98	0.72	5	92	3	0	99	1
8	241	115.699	0.14	2.401	2.685	1.20	0.62	1	98	1	0	100	0
9	112	115.164	-0.38	4.047	2.734	0.68	-0.95	10	80	10	4	88	8
10	122	115.597	0.04	3.846	2.816	0.73	0.12	7	88	5	2	94	4
11	132	115.139	-0.41	5.108	4.371	0.85	-0.80	3	92	5	1	97	2
12	142	115.272	-0.27	4.598	4.782	1.04	-0.60	1	97	2	0	99	1
13	212	115.724	0.17	2.965	2.022	0.68	0.59	12	81	7	8	88	4
14	222	114.890	-0.65	2.614	2.092	0.74	-2.34	5	87	8	2	93	5
15	232	115.639	0.09	3.384	3.084	0.91	0.27	8	89	3	3	96	1
16	242	115.674	0.12	2.972	3.338	1.12	0.42	3	95	2	0	100	0
17	113	114.989	-0.58	3.873	2.756	0.71	-1.50	6	85	9	3	92	5
18	123	115.349	-0.20	3.476	2.554	0.73	-0.58	7	87	8	1	95	4
19	133	115.184	-0.36	4.193	3.892	0.93	-0.87	4	92	4	2	95	3
20	143	115.338	-0.21	3.690	3.905	1.06	-0.57	1	96	3	0	98	2
21	213	115.771	0.22	2.934	2.046	0.70	0.76	11	83	6	7	90	3
22	223	114.870	-0.67	2.614	2.008	0.77	-2.60	6	84	10	2	94	4
23	233	115.848	0.29	2.943	2.773	0.94	1.01	7	90	3	1	98	1
24	243	115.698	0.14	2.505	2.889	1.15	0.59	1	98	1	0	100	0
25	114	115.252	-0.29	3.934	2.693	0.68	-0.76	11	79	10	6	89	5
26	124	115.684	0.13	3.725	2.764	0.74	0.36	8	87	5	3	94	3
27	134	115.262	-0.28	4.945	4.300	0.87	-0.58	3	93	4	1	97	2
28	144	115.380	-0.16	4.451	4.883	1.08	-0.38	1	97	2	0	99	1
29	214	115.812	0.26	3.021	2.020	0.67	0.87	13	77	10	9	87	4
30	224	114.856	-0.89	2.737	2.047	0.75	-2.53	5	87	8	2	93	5
31	234	115.783	0.23	3.476	3.095	0.89	0.67	9	88	3	2	97	1
32	244	115.648	0.09	2.970	3.279	1.10	0.33	2	96	2	0	100	0
33	115	115.189	-0.36	4.056	2.736	0.67	-0.89	11	78	11	4	88	8
34	125	115.590	0.04	3.862	2.813	0.73	0.11	8	87	5	2	94	4
35	135	115.201	-0.34	5.151	4.378	0.85	-0.88	3	92	5	1	97	2
36	145	115.305	-0.24	4.645	4.784	1.03	-0.53	1	97	2	0	99	1
37	215	116.073	0.52	3.083	2.032	0.66	1.70	15	78	7	10	85	5
38	225	115.022	-0.52	2.817	2.092	0.74	-1.87	5	88	7	2	94	4
39	235	116.048	0.49	3.561	3.124	0.87	1.39	9	88	3	3	97	0
40	245	115.833	0.28	3.070	3.367	1.10	0.92	3	95	2	0	100	0
41	116	115.046	-0.50	3.765	2.721	0.72	-1.34	6	85	9	5	92	3
42	126	115.397	-0.15	3.428	2.545	0.74	-0.45	7	87	5	1	95	4
43	136	115.251	-0.29	4.133	3.849	0.93	-0.72	4	92	4	2	95	3
44	146	115.373	-0.17	3.739	3.899	1.04	-0.47	2	95	3	0	98	2
45	216	115.836	0.28	2.992	2.018	0.67	0.96	14	79	7	8	88	4
46	226	114.844	-0.70	2.574	1.997	0.78	-2.74	5	85	10	2	84	4
47	236	115.816	0.26	2.921	2.737	0.94	0.91	7	90	3	1	98	1
48	246	115.643	0.09	2.467	2.865	1.16	0.38	1	98	1	0	100	0
49	117	114.998	-0.55	3.884	2.757	0.71	-1.42	6	85	9	3	92	5
50	127	115.345	-0.20	3.483	2.553	0.73	-0.59	7	85	7	1	95	4
51	137	115.225	-0.32	4.217	3.895	0.92	-0.77	4	92	4	2	95	3
52	147	115.332	-0.21	3.688	3.905	1.06	-0.59	1	96	3	0	98	2
53	217	116.094	0.54	3.047	2.050	0.67	1.79	14	80	6	8	89	3
54	227	114.939	-0.61	2.609	2.006	0.77	-2.34	6	84	10	2	94	4
55	237	115.990	0.44	3.013	2.788	0.92	1.46	9	89	2	2	97	1
56	247	115.742	0.19	2.501	2.867	1.15	0.77	1	98	1	0	100	0
57	118	115.651	0.10	3.572	2.508	0.68	0.28	10	83	7	5	91	4
58	128	115.885	0.33	3.602	2.438	0.68	0.83	13	83	4	5	91	4
59	138	115.706	0.15	3.858	3.542	0.92	0.41	4	93	3	2	96	2
60	148	115.725	0.17	3.688	3.753	1.02	0.48	2	95	3	0	98	2
61	218	115.457	-0.09	2.873	1.883	0.65	-0.32	12	82	6	7	89	4
62	228	114.683	-0.86	2.669	1.878	0.70	-3.24	3	86	11	2	91	7
63	238	115.417	-0.13	2.931	2.652	0.90	-0.45	7	90	3	2	96	2
64	248	115.490	-0.06	2.584	2.760	1.07	-0.23	3	94	3	0	99	1
65	119	115.583	0.03	3.753	2.512	0.67	0.09	12	79	9	5	91	4
66	129	115.783	0.23	3.724	2.446	0.66	0.63	14	81	5	5	91	4
67	139	115.651	0.10	3.959	3.546	0.90	0.26	5	92	3	2	96	2
68	149	115.646	0.09	3.811	3.768	0.99	0.25	1	96	3	0	98	2
69	219	115.710	0.16	2.934	1.872	0.64	0.55	15	79	5	6	88	4
70	229	114.843	-0.70	2.744	1.890	0.69	-2.57	6	85	8	3	92	5
71	239	115.682	0.13	2.990	2.638	0.88	0.44	8	89	3	3	95	2
72	249	115.665	0.11	2.649	2.778	1.05	0.44	4	93	3	0	99	1

TABLE 2: The statistics \bar{z} , $(\bar{z}-\mu)$, $\sqrt{S_{ZZ}}$, $\sqrt{\bar{V}}$, the ratio $\sqrt{\bar{V}/S_{ZZ}}$, t and the number of times μ fell below, within or above the 95% and 99% confidence intervals, as calculated by procedures 1, 2, 3, 4, 5, 6, 7, 8, and 9 for population 2; Sampling method: 15 percent of trees selected from 30, 20 and 10 plots of stratum 1, 2 and 3 respectively.

estimation model/ procedure	\bar{z}	$(\bar{z}-\mu)$	$\sqrt{S_{ZZ}}$	$\sqrt{\bar{V}}$	$\sqrt{\bar{V}/S_{ZZ}}$	t	number of times						
							95%			99%			
							b	w	a	b	w	a	
1	111	117.689	3.83	3.746	2.722	0.73	10.23	36	62	2	23	77	0
2	121	117.859	4.00	3.659	2.876	0.78	10.85	36	63	1	20	80	0
3	131	117.707	3.85	3.861	3.965	1.03	9.98	18	82	0	6	94	0
4	141	117.783	3.92	3.857	4.425	1.15	10.19	12	88	0	2	98	0
5	211	118.553	4.69	2.538	2.057	0.81	18.52	64	36	0	41	59	0
6	221	116.401	2.54	2.842	2.643	0.93	8.95	17	82	1	4	96	0
7	231	118.477	4.62	2.722	3.029	1.11	16.99	36	64	0	15	85	0
8	241	117.571	3.71	2.805	3.708	1.32	13.25	10	90	0	1	99	0
9	112	113.469	-0.38	4.047	2.734	0.68	-0.95	10	80	10	4	88	8
10	122	113.901	0.04	3.846	2.816	0.73	0.12	7	88	5	2	94	4
11	132	113.444	-0.41	5.108	4.371	0.86	-0.80	3	92	5	1	97	2
12	142	113.577	-0.27	4.598	4.782	1.04	-0.60	1	97	2	0	99	1
13	212	114.028	0.17	2.965	2.022	0.68	0.59	12	81	7	8	88	4
14	222	113.195	-0.65	2.814	2.092	0.74	-2.34	5	87	8	2	93	5
15	232	113.944	0.09	3.384	3.084	0.91	0.27	8	89	3	3	96	1
16	242	113.978	0.12	2.972	3.338	1.12	0.42	3	95	2	0	100	0
17	113	113.430	-0.42	3.937	2.861	0.73	-1.08	6	85	9	3	92	5
18	123	113.872	0.01	3.612	2.728	0.76	0.05	9	86	5	1	96	3
19	133	114.338	0.48	4.798	4.030	0.84	1.01	9	86	5	2	96	2
20	143	114.168	0.31	4.270	4.222	0.99	0.74	4	94	2	0	99	1
21	213	114.090	0.23	2.998	2.062	0.69	0.79	12	82	6	7	90	3
22	223	113.180	-0.67	2.736	2.212	0.81	-2.46	5	88	7	2	96	2
23	233	114.197	0.34	3.335	2.969	0.89	1.03	7	90	3	2	96	2
24	243	113.779	-0.07	2.853	3.165	1.11	-0.26	1	98	1	0	100	0
25	114	114.527	0.67	3.934	2.693	0.68	1.71	11	82	7	6	90	4
26	124	114.959	1.10	3.725	2.764	0.74	2.97	12	84	4	5	93	2
27	134	114.536	0.68	4.945	4.300	0.87	1.38	4	93	3	2	96	2
28	144	114.655	0.80	4.451	4.563	1.05	1.80	2	96	2	0	99	1
29	214	115.087	1.23	3.021	2.020	0.67	4.08	21	75	4	12	85	3
30	224	114.131	0.27	2.737	2.047	0.75	1.01	10	84	8	4	94	2
31	234	115.058	1.20	3.476	3.095	0.89	3.46	12	87	1	5	95	0
32	244	114.923	1.06	2.970	3.279	1.10	3.60	5	94	1	1	99	0
33	115	113.273	-0.58	4.056	2.736	0.67	-1.43	8	81	11	4	88	8
34	125	113.674	-0.18	3.882	2.813	0.73	-0.47	7	87	6	2	93	5
35	135	113.284	-0.56	5.151	4.378	0.85	-1.11	3	91	6	1	97	2
36	145	113.389	-0.46	4.645	4.784	1.03	-1.00	1	96	3	0	99	1
37	215	114.157	0.30	3.083	2.032	0.66	0.98	12	80	8	8	87	5
38	225	113.106	-0.74	2.817	2.092	0.74	-2.65	5	88	7	2	94	4
39	235	114.132	0.27	3.581	3.124	0.87	0.78	7	90	3	2	96	2
40	245	113.917	0.06	3.070	3.367	1.10	0.21	3	95	2	0	100	0
41	116	114.486	0.63	3.828	2.813	0.73	1.65	8	87	5	5	92	3
42	126	114.882	1.02	3.525	2.691	0.76	2.92	12	85	3	4	94	2
43	136	115.308	1.45	4.556	3.974	0.87	3.19	9	88	3	5	93	2
44	146	114.996	1.14	4.124	4.183	1.01	2.77	6	92	2	0	99	1
45	216	115.141	1.28	3.034	2.031	0.67	4.24	21	75	4	14	83	3
46	226	114.098	0.24	2.663	2.170	0.81	0.92	6	90	4	3	95	2
47	236	115.141	1.28	3.368	2.926	0.87	3.82	15	83	2	4	94	2
48	246	114.435	0.58	2.808	3.119	1.11	2.07	2	97	1	0	100	0
49	117	113.236	-0.61	3.947	2.862	0.72	-1.57	6	85	9	2	92	6
50	127	113.689	-0.16	3.624	2.724	0.75	-0.46	7	86	7	1	95	4
51	137	114.235	0.38	4.845	4.032	0.83	0.79	9	86	5	1	97	2
52	147	114.064	0.21	4.297	4.220	0.98	0.49	3	94	3	0	99	1
53	217	114.217	0.36	3.103	2.065	0.67	1.17	13	80	7	8	89	3
54	227	113.050	-0.80	2.728	2.205	0.81	-2.95	5	88	7	2	96	2
55	237	114.269	0.41	3.485	2.976	0.85	1.19	7	89	4	3	95	2
56	247	113.714	-0.14	2.837	3.158	1.11	-0.49	1	97	2	0	100	0
57	118	115.172	1.31	3.694	2.691	0.73	3.57	18	79	3	4	94	2
58	128	115.485	1.63	3.635	2.921	0.80	4.49	13	84	3	5	93	2
59	138	115.257	1.40	3.822	3.998	1.05	3.67	4	94	2	1	98	1
60	148	115.438	1.58	3.787	4.488	1.19	4.18	0	98	2	0	100	0
61	218	114.985	1.13	2.830	2.082	0.74	4.00	18	80	4	10	88	2
62	228	112.828	-1.02	2.843	2.751	0.97	-3.61	1	96	3	0	97	3
63	238	114.981	1.12	2.913	3.157	1.08	3.87	6	92	2	2	98	0
64	248	114.070	0.21	2.870	3.860	1.34	0.75	0	99	1	0	100	0
65	119	113.958	0.10	3.746	2.684	0.72	0.28	5	91	4	2	95	3
66	129	114.219	0.36	3.717	2.947	0.79	0.98	8	89	3	1	96	3
67	139	114.094	0.24	3.910	3.973	1.02	0.62	2	96	2	1	97	2
68	149	114.241	0.38	3.872	4.494	1.16	1.00	0	98	2	0	98	2
69	219	114.103	0.24	2.861	2.064	0.72	0.87	13	83	4	5	92	3
70	229	111.685	-2.16	2.969	2.811	0.95	-7.30	0	89	11	0	97	2
71	239	114.147	0.29	2.956	3.134	1.06	0.99	3	95	2	1	97	2
72	249	112.988	-0.86	3.003	3.911	1.30	-2.88	0	99	1	0	100	0

TABLE 3: The statistics \bar{z} , $(\bar{z}-\mu)$, $\sqrt{S_{zz}}$, \sqrt{v} , the ratio \sqrt{v}/S_{zz} , t and the number of times μ fell below, within or above the 95% and 99% confidence intervals, as calculated by procedures 1, 2, 3, 4, 5, 6, 7, 8 and 9 for population 3; Sampling method : 15 percent of trees selected from 30, 20 and 10 plots of stratum 1, 2 and 3 respectively.

estimation model/ procedure	\bar{z}	$(\bar{z}-\mu)$	$\sqrt{S_{zz}}$	\sqrt{v}	\sqrt{v}/S_{zz}	t	number of times						
							95%			99%			
							b	w	a	b	w	a	
1	111	126.759	7.12	3.508	2.758	0.79	20.31	68	32	0	52	48	0
2	121	126.903	7.26	3.474	3.298	0.95	20.92	56	44	0	33	67	0
3	131	126.790	7.15	3.738	4.303	1.15	19.14	36	64	0	20	80	0
4	141	126.865	7.23	3.740	5.182	1.39	19.33	20	80	0	8	92	0
5	211	127.414	7.78	2.667	2.252	0.84	29.18	88	12	0	76	24	0
6	221	125.201	5.56	3.187	3.441	1.08	17.47	35	65	0	17	83	0
7	231	127.218	7.58	2.909	3.825	1.31	26.07	49	51	0	27	73	0
8	241	125.966	6.33	3.387	5.153	1.52	18.70	14	86	0	2	98	0
9	112	119.249	-0.38	4.047	2.734	0.68	-0.95	10	80	10	4	88	8
10	122	119.681	0.04	3.846	2.816	0.73	0.12	7	88	5	2	94	4
11	132	119.224	-0.41	5.108	4.371	0.86	-0.80	3	92	5	1	97	2
12	142	119.357	-0.27	4.598	4.782	1.04	-0.60	1	97	2	0	99	1
13	212	119.809	0.17	2.965	2.022	0.68	0.59	12	81	7	8	88	4
14	222	118.975	-0.65	2.814	2.092	0.74	-2.34	5	87	8	2	93	5
15	232	119.724	0.09	3.384	3.084	0.91	0.27	8	89	3	3	96	1
16	242	119.759	0.12	2.972	3.338	1.12	0.42	3	95	2	0	100	0
17	113	119.263	-0.37	3.985	2.854	0.72	-0.93	8	83	9	3	92	5
18	123	120.777	1.14	3.736	2.610	0.70	3.06	17	79	4	8	90	2
19	133	119.637	0.00	4.587	4.242	0.92	0.01	6	88	6	3	94	3
20	143	120.819	1.18	3.886	4.120	1.06	3.05	0	98	2	0	99	1
21	213	119.830	0.19	2.939	2.062	0.70	0.67	11	83	6	6	91	3
22	223	119.633	0.00	3.028	2.111	0.70	0.00	9	86	5	3	95	2
23	233	119.906	0.27	3.186	2.959	0.93	0.85	4	93	3	1	98	1
24	243	120.994	1.36	2.581	3.016	1.17	5.27	5	95	0	0	100	0
25	114	118.781	-0.85	3.934	2.693	0.68	-2.17	7	82	11	4	90	6
26	124	119.213	-0.42	3.725	2.764	0.74	-1.13	6	88	6	2	93	5
27	134	118.790	-0.84	4.845	4.300	0.87	-1.71	3	93	4	1	97	2
28	144	118.909	-0.72	4.451	4.683	1.05	-1.63	1	95	4	0	99	1
29	214	119.341	-0.29	3.021	2.020	0.67	-0.97	11	77	12	7	89	4
30	224	118.385	-1.24	2.737	2.047	0.75	-4.56	4	86	10	0	94	6
31	234	119.312	-0.32	3.476	3.095	0.89	-0.93	4	93	3	0	99	1
32	244	119.177	-0.45	2.970	3.279	1.10	-1.54	2	96	2	0	100	0
33	115	119.319	-0.31	4.056	2.736	0.67	-0.78	11	79	10	4	88	8
34	125	119.719	0.08	3.862	2.813	0.73	0.22	8	87	5	3	93	4
35	135	119.330	-0.30	5.151	4.378	0.85	-0.59	3	92	5	1	97	2
36	145	119.435	-0.19	4.645	4.784	1.03	-0.43	1	97	2	0	99	1
37	215	120.203	0.56	3.083	2.032	0.66	1.84	15	79	6	10	85	5
38	225	119.152	-0.48	2.817	2.092	0.74	-1.71	5	88	7	3	93	4
39	235	120.178	0.54	3.581	3.124	0.87	1.52	9	89	2	3	97	0
40	245	119.963	0.32	3.070	3.367	1.10	1.07	3	95	2	0	100	0
41	116	118.796	-0.83	3.882	2.807	0.72	-2.16	5	85	10	4	92	4
42	126	120.185	0.95	3.632	2.592	0.71	1.52	10	84	6	4	93	3
43	136	119.200	-0.43	4.492	4.162	0.93	-0.97	5	90	5	3	94	3
44	146	120.236	0.60	3.841	4.088	1.06	1.57	0	98	2	0	99	1
45	216	119.320	-0.31	2.999	2.032	0.68	-1.05	9	81	10	6	90	4
46	226	118.968	-0.66	2.812	2.082	0.72	-2.29	6	86	8	2	94	4
47	236	119.199	-0.43	3.161	2.903	0.92	-1.38	2	94	4	1	97	2
48	246	120.213	0.57	2.497	2.988	1.20	2.32	1	98	1	0	100	0
49	117	119.332	-0.30	3.995	2.855	0.71	-0.76	8	83	9	4	91	5
50	127	120.787	1.15	3.740	2.609	0.70	3.08	17	79	4	8	90	2
51	137	119.710	0.07	4.610	4.245	0.92	0.16	6	88	6	3	94	3
52	147	120.817	1.18	3.892	4.121	1.06	3.04	0	98	2	0	99	1
53	217	120.186	0.55	3.061	2.067	0.68	1.80	14	80	6	8	89	3
54	227	119.735	0.10	2.996	2.106	0.70	0.34	9	86	5	3	95	2
55	237	120.103	0.46	3.252	2.953	0.91	1.44	6	92	2	2	97	1
56	247	121.054	1.42	2.557	3.014	1.16	5.55	5	95	0	0	100	0
57	118	119.261	-0.37	3.791	2.741	0.72	-0.99	8	85	7	4	93	3
58	128	119.466	-0.16	3.773	3.231	0.86	-0.45	5	91	4	0	98	2
59	138	119.493	-0.14	3.996	4.305	1.08	-0.35	3	94	3	1	98	1
60	148	119.539	-0.09	3.942	5.105	1.30	-0.24	0	98	2	0	99	1
61	218	119.012	-0.62	3.020	2.254	0.75	-2.06	7	87	6	2	94	4
62	228	116.748	-2.88	3.132	3.358	1.07	-8.22	0	92	8	0	96	4
63	238	119.082	-0.55	3.119	3.822	1.23	-1.77	2	96	2	0	99	1
64	248	117.801	-1.83	3.268	5.032	1.54	-5.61	0	100	0	0	100	0
65	119	119.802	0.16	3.873	2.784	0.71	0.43	11	83	6	4	93	3
66	129	119.964	0.33	3.892	3.290	0.85	0.85	7	90	3	1	97	2
67	139	120.059	0.42	4.103	4.365	1.06	1.04	3	94	3	1	98	1
68	149	120.085	0.45	4.077	5.203	1.28	1.11	0	98	2	0	99	1
69	219	119.850	0.21	3.103	2.273	0.73	0.70	13	82	5	5	93	2
70	229	117.313	-2.32	3.284	3.441	1.05	-7.07	0	93	7	0	97	3
71	239	119.905	0.27	3.238	3.893	1.20	0.84	2	96	2	1	98	1
72	249	118.358	-1.27	3.455	5.154	1.49	-3.69	0	100	0	0	100	0

TABLE 4: The statistics \bar{z} , $(\bar{z}-\mu)$, $\sqrt{s_{zz}}$, \sqrt{v} , the ratio \sqrt{v}/s_{zz} , t and the number of times μ fell below, within or above the 95% and 99% confidence intervals, as calculated by procedures 1, 2, 3, 4, 5, 6, 7, 8 and 9 for population 4; Sampling method: 15 percent of trees selected from 30, 20 and 10 plots of stratum 1, 2 and 3 respectively.

estimation model/ procedure	\bar{z}	$(\bar{z}-\mu)$	$\sqrt{s_{zz}}$	\sqrt{v}	\sqrt{v}/s_{zz}	t	number of times						
							95%			99%			
							b	w	a	b	w	a	
1	111	131.543	-6.00	3.577	2.747	0.77	-16.78	0	46	54	0	58	42
2	121	131.730	-5.81	3.405	3.107	0.91	-17.08	0	53	47	0	74	26
3	131	131.850	-5.89	3.888	4.251	1.09	-15.17	1	75	24	0	91	9
4	141	131.738	-5.80	3.699	5.003	1.35	-15.70	0	83	17	0	94	6
5	211	132.291	-5.25	2.874	2.202	0.77	-18.29	0	33	67	0	54	46
6	221	131.374	-6.17	3.076	2.994	0.97	-20.07	0	46	54	0	69	31
7	231	132.231	-5.31	3.094	3.637	1.18	-17.18	0	72	28	0	90	10
8	241	132.168	-5.37	3.019	4.621	1.53	-17.81	0	87	13	0	96	4
9	112	137.162	-0.38	4.047	2.734	0.68	-0.95	10	80	10	4	88	8
10	122	137.594	0.04	3.846	2.816	0.73	0.12	7	88	5	2	94	4
11	132	137.137	-0.41	5.108	4.371	0.86	-0.80	3	92	5	1	97	2
12	142	137.270	-0.27	4.598	4.782	1.04	-0.60	1	97	2	0	99	1
13	212	137.722	0.17	2.865	2.022	0.68	0.59	12	81	7	8	88	4
14	222	136.888	-0.65	2.814	2.092	0.74	-2.34	5	87	8	2	93	5
15	232	137.636	0.09	3.384	3.084	0.91	0.27	8	89	3	3	96	1
16	242	137.671	0.12	2.972	3.338	1.12	0.42	3	95	2	0	100	0
17	113	137.125	-0.42	3.844	2.855	0.72	-1.07	6	85	9	3	92	5
18	123	136.604	-0.94	3.731	2.625	0.70	-2.53	6	80	14	1	94	5
19	133	137.135	-0.41	10.377	6.385	0.62	-0.40	6	86	8	3	94	3
20	143	136.740	-0.80	6.686	5.178	0.77	-1.21	1	95	4	0	99	1
21	213	137.763	0.21	2.979	2.063	0.69	0.73	12	82	6	7	90	3
22	223	136.447	-1.10	2.828	2.132	0.75	-3.89	5	86	9	1	94	5
23	233	137.805	0.25	3.338	2.976	0.89	0.77	7	90	3	2	97	1
24	243	136.858	-0.68	2.651	3.021	1.14	-2.60	0	98	2	0	99	1
25	114	135.278	-2.26	3.934	2.693	0.68	-5.76	5	73	22	1	88	11
26	124	135.711	-1.83	3.725	2.764	0.74	-4.93	3	84	13	0	95	5
27	134	135.288	-2.25	4.845	4.300	0.87	-4.57	3	87	10	0	97	3
28	144	135.407	-2.13	4.451	4.683	1.05	-4.81	0	94	6	0	98	2
29	214	135.839	-1.70	3.021	2.020	0.67	-5.65	2	78	20	0	86	12
30	224	134.883	-2.66	2.737	2.047	0.75	-9.73	0	70	30	0	85	15
31	234	135.810	-1.73	3.476	3.095	0.89	-5.00	0	91	9	0	97	3
32	244	135.675	-1.87	2.870	3.279	1.10	-6.30	0	94	6	0	98	2
33	115	137.230	-0.31	4.056	2.736	0.67	-0.78	11	79	10	4	88	8
34	125	137.630	0.08	3.862	2.813	0.73	0.22	8	87	5	3	93	4
35	135	137.241	-0.30	5.151	4.378	0.85	-0.59	3	92	5	1	97	2
36	145	137.346	-0.20	4.645	4.784	1.03	-0.43	1	97	2	0	99	1
37	215	138.114	0.56	3.083	2.032	0.66	1.84	15	79	6	10	85	5
38	225	137.063	-0.48	2.817	2.092	0.74	-1.72	5	88	7	3	93	4
39	235	138.089	0.54	3.581	3.124	0.87	1.51	9	88	3	3	97	0
40	245	137.874	0.32	3.070	3.367	1.10	1.06	3	95	2	0	100	0
41	116	135.248	-2.29	3.837	2.808	0.73	-5.99	4	77	19	0	90	10
42	126	134.907	-2.63	3.614	2.606	0.72	-7.30	1	78	21	0	86	14
43	136	135.315	-2.23	8.781	6.094	0.62	-2.28	3	83	14	1	93	6
44	146	135.041	-2.50	6.396	5.039	0.79	-3.92	1	89	10	0	97	3
45	216	135.842	-1.70	3.027	2.033	0.67	-5.63	2	76	22	1	88	11
46	226	134.514	-3.03	2.689	2.101	0.78	-11.24	0	68	32	0	87	13
47	236	135.726	-1.82	3.390	2.931	0.86	-5.37	1	90	9	0	97	3
48	246	134.983	-2.58	2.572	2.996	1.17	-10.05	0	89	11	0	96	2
49	117	137.190	-0.35	3.953	2.856	0.72	-0.90	6	85	9	4	91	5
50	127	136.609	-0.93	3.738	2.623	0.70	-2.51	6	80	14	1	94	5
51	137	137.197	-0.35	10.386	6.381	0.61	-0.34	6	86	8	3	94	3
52	147	136.740	-0.80	6.651	5.176	0.78	-1.21	1	95	4	0	99	1
53	217	138.119	0.57	3.084	2.068	0.67	1.86	14	80	6	9	88	3
54	227	136.549	-0.99	2.845	2.126	0.75	-3.51	5	86	9	1	94	5
55	237	138.021	0.47	3.479	2.982	0.86	1.36	10	88	2	4	95	1
56	247	136.928	-0.61	2.680	3.025	1.13	-2.31	0	98	2	0	99	1
57	118	135.588	-1.97	3.819	2.793	0.73	-5.18	2	79	19	1	88	11
58	128	135.929	-1.61	3.744	3.236	0.86	-4.32	1	90	9	0	96	4
59	138	135.560	-1.98	4.217	4.453	1.06	-4.71	1	91	8	1	96	3
60	148	135.670	-1.87	4.105	5.283	1.29	-4.57	1	95	4	0	99	1
61	218	135.404	-2.14	3.036	2.291	0.75	-7.06	2	74	24	1	89	10
62	228	135.056	-2.49	3.425	3.132	0.91	-7.27	0	88	12	0	95	5
63	238	135.301	-2.24	3.201	3.871	1.21	-7.02	0	94	6	0	98	2
64	248	135.810	-1.73	3.165	4.844	1.53	-5.49	0	98	2	0	100	0
65	119	137.477	-0.07	3.893	2.778	0.71	-0.18	9	81	10	2	92	6
66	129	137.820	0.27	3.864	3.189	0.83	0.71	7	89	4	1	97	2
67	139	137.426	-0.12	4.274	4.401	1.03	-0.28	2	93	5	1	97	2
68	149	137.519	-0.02	4.186	5.207	1.24	-0.07	1	97	2	0	99	1
69	219	137.620	0.07	3.091	2.252	0.73	0.24	10	84	6	3	93	4
70	229	137.267	-0.28	3.431	3.052	0.89	-0.82	2	94	4	0	97	3
71	239	137.452	-0.09	3.249	3.780	1.16	-0.29	1	97	2	0	99	1
72	249	138.005	0.45	3.158	4.721	1.49	1.45	0	100	0	0	100	0

Table 5: Precision $\sqrt{S_{zz}}$ of estimates z by (i) form of regression function, (ii) least squares estimation approach (iii) procedure for taking into account the stratum effect and (iv) population.

==POPULATION 1==										
Least squares method	Regression form	Procedure number								
		1	2	3	4	5	6	7	8	9
OLS	1	3.57	4.05	3.87	3.93	4.06	3.76	3.88	3.67	3.75
OWLS	1	3.43	3.85	3.48	3.73	3.86	3.43	3.48	3.60	3.72
MLS	1	3.72	5.11	4.19	4.94	5.15	4.13	4.22	3.86	3.96
MWLS	1	3.57	4.60	3.69	4.45	4.65	3.74	3.69	3.69	3.81
OLS	2	2.57	2.96	2.93	3.02	3.08	2.99	3.05	2.87	2.93
OWLS	2	2.46	2.81	2.61	2.74	2.82	2.57	2.61	2.67	2.74
MLS	2	2.71	3.38	2.94	3.48	3.58	2.92	3.01	2.93	2.99
MWLS	2	2.40	2.97	2.50	2.97	3.07	2.47	2.50	2.58	2.65
==POPULATION 2==										
OLS	1	3.75	4.05	3.94	3.93	4.06	3.83	3.95	3.69	3.75
OWLS	1	3.66	3.85	3.61	3.73	3.86	3.53	3.62	3.64	3.72
MLS	1	3.86	5.11	4.80	4.94	5.15	4.56	4.84	3.82	3.91
MWLS	1	3.86	4.60	4.27	4.45	4.65	4.12	4.30	3.79	3.87
OLS	2	2.54	2.96	3.00	3.02	3.08	3.03	3.10	2.83	2.86
OWLS	2	2.84	2.81	2.74	2.74	2.82	2.66	2.73	2.84	2.97
MLS	2	2.72	3.38	3.34	3.48	3.58	3.37	3.49	2.91	2.96
MWLS	2	2.80	2.97	2.85	2.97	3.07	2.81	2.84	2.87	3.00
==POPULATION 3==										
OLS	1	3.51	4.05	3.99	2.94	4.06	3.88	4.00	3.79	3.87
OWLS	1	3.47	3.85	3.74	3.03	3.86	3.63	3.74	3.77	3.89
MLS	1	3.74	5.11	4.59	3.19	5.15	4.49	4.61	4.00	4.10
MWLS	1	3.74	4.60	3.89	2.58	4.65	3.84	3.89	3.94	4.08
OLS	2	2.67	2.96	2.94	3.93	3.08	3.00	3.06	3.02	3.10
OWLS	2	3.19	2.81	3.03	3.73	2.81	2.91	3.00	3.13	3.28
MLS	2	2.91	3.38	3.19	4.94	3.58	3.16	3.25	3.12	3.24
MWLS	2	3.39	2.97	2.58	4.45	3.07	2.50	2.56	3.27	3.46
==POPULATION 4==										
OLS	1	3.58	4.05	3.94	3.93	4.06	3.84	3.95	3.82	3.89
OWLS	1	3.40	3.85	3.73	3.73	3.86	3.61	3.74	3.74	3.86
MLS	1	3.89	5.11	10.38	4.94	5.15	9.78	10.39	4.22	4.27
MWLS	1	3.70	4.60	6.69	4.45	4.65	6.40	6.65	4.10	4.19
OLS	2	2.87	2.96	2.98	3.02	3.08	3.03	3.08	3.04	3.09
OWLS	2	3.08	2.81	2.83	2.74	2.82	2.70	2.84	3.42	3.43
MLS	2	3.09	3.38	3.34	3.48	3.58	3.39	3.48	3.20	3.25
MWLS	2	3.02	2.97	2.65	2.97	3.07	2.57	2.68	3.16	3.16

weighted regressions with weights based on area as a measure of stratum size is not a good procedure and (ii) using weighted regression with weights based on number of trees as a measure of stratum size may lead, although not always, to unbiased results.

To summarize the above discussion it appears that (i) Procedure 2 is always unbiased, (ii) Procedure 3 is almost always unbiased and (iii) if single biomass regressions applicable to all strata are desired, the individual stratum data (Procedures 8 and 9) or regressions (Procedures 4, 5, 6 and 7) should be weighted by number of trees, not area of the strata.

Analysis of Precision $\sqrt{S_{zz}}$ and Accuracy of z

To analyze the precision of z we could have used Tables 1, 2, 3 and 4. The analysis is facilitated, however, when all the estimates $\sqrt{S_{zz}}$ of the precision of z are rearranged in a separate Table 5 by least squares estimation approach, regression function form, procedure number and population.

Because Procedure 2 is the standard against which the other procedures are usually compared, let us start with its analysis. An ocular scan of the results in Table 5 shows that, as expected, (i) the weighted least squares estimators are slightly more precise than the least squares ones, (ii) the modified estimators are less precise than the ordinary ones and (iii) the estimators based on diameter and height (regression form 2) are much more precise than the ones based on diameter alone (regression form 1). These conclusions seem to hold approximately true for the other procedures as well, although sometimes, the weighted least squares estimators turn out to be less precise, than the corresponding least squares.

A look at the precision of the estimators calculated by Procedure 1 (using a single biomass regression function for all strata) shows that it is higher (that is, the value of $\sqrt{S_{zz}}$ is lower) than that of the estimators calculated by Procedure 2 (using a different biomass regression for each stratum). This seems surprising until we realize that the estimates of the single regression function are based on a much larger number of sample trees and, thus are much more precise

than the estimates of the three individual regressions of Procedure 2. On the other hand the estimators of Procedure 1 are less accurate. If we measure the accuracy by the mean square error expression $(S_{zz} + (\bar{z} - \mu)^2)$, and because the bias by Procedure 2 is generally smaller, the mean square error of the estimators by Procedure 1 turns out to be higher in value, and, thus, the estimators turn out to have a lower accuracy.

The estimates based on Procedure 3 (where parallel or identical regressions are used whenever warranted by the results of the significance tests) are generally more precise than those based on Procedure 2. Occasionally, however, they are much less precise. For example, the precision of the MLS estimators using regression form 1 of population 4 is extremely low. This inconsistency is hard to explain. Furthermore, the bias of the estimators by Procedure 3 is, in some instances, relatively large and significantly different from zero. Consequently, the estimators calculated by Procedure 2 seem to be more consistent and, on the average more accurate than those calculated by Procedure 3.

The precision of the estimators derived by Procedures 4 and 5 is very close to that of Procedure 2, and the precision of the estimators derived by Procedures 6 and 7 is very close to that of Procedure 3. This seems to imply that it does not matter whether we use separate regressions, one for each stratum, or a single regression for all strata, the weighted average of these separate regressions. However, because (i) there is always a real possibility of introducing a large bias when using Procedures 4 and 6 and (ii) it is questionable whether good estimators of the number of trees by strata (used as weights) can ever be obtained, Procedures 2 and 3 are to be preferred to Procedures 4, 5, 6 and 7.

Finally, the estimates based on Procedures 8 and 9 seem to have relatively good precision for population 1. For the other three populations the results seem to be mixed. As the estimates based on Procedure 8 are biased and the ones based on Procedure 9 require the knowledge of the total number of trees contained in each stratum, there seems to be no advantage in using them in preference to those generated by Procedure 2.

Analysis of the Estimated Precision \sqrt{V} of z

The sample based estimator of the precision of z is the sample variance V , or its square root, the standard error \sqrt{V} . Because S_{zz} is an unbiased estimator of the variance of z (which is independent of the assumptions of the estimation model), the ratio V/S_{zz} or its square root $\sqrt{V/S_{zz}}$ can be used as a measure of the goodness of the estimator V . We shall use the square root. And to facilitate the analysis, the values of the ratios $\sqrt{V/S_{zz}}$ of Tables 1, 2, 3 and 4 were rearranged in Table 6 by population, regression function form, procedure number and least squares estimation approach.

As the reader can verify by an ocular analy-

sis of the results listed in Table 6, the ratio $\sqrt{V/S_{zz}}$ is consistently much lower in value than 1 for all ordinary least and weighted least squares estimators, with slightly lower values for the least than weighted least squares. This shows that V is a gross underestimate of the variance of z . On the average, the value of the ratio is about .70 for the OLS and about .78 for the OWLS estimators of all procedures. The statistic V of the modified least squares is also an underestimate of the variance of z ; on the average, the value of the ratio is about .90. On the other hand, the ratio for the modified weighted least squares estimators is over 1; this means that the variance of the MWLS estimators is somewhat overestimated by V .

Analysis of the Confidence Intervals of μ

The confidence intervals are a good expression of the combined effect of precision and bias; and for the layman, a much more intuitive and meaningful measure of the validity and accuracy of the estimators. A look at the number of times the 95 and 99 percent confidence intervals fall above, over or below μ leads to the conclusion that the confidence intervals are (i) badly off when the estimation procedures are based on the ordinary least and weighted least squares and (ii) relatively all right when the estimates are calculated by the modified procedures. It should be noted here that the proportion of times the confidence interval statements were found to be right is significantly different from the expected 95 or 99 times for the ordinary but not significantly different for the modified techniques.

Analysis of Results - Part 2

In the previous section we have made a detailed analysis of the results obtained when the stratified sample was of a specific size and allocation. The objective of the present section is to analyze other sample sizes and allocations and see whether the main calculations reached in the previous section apply in these cases as well; or, to see what changes, if any, should we make to our conclusions when the percentage p of trees selected from a sample cluster changes from the value $p=15$ used, and the number of clusters selected per stratum 1, 2, and 3 is different from that of 30, 20 and 10 respectively of the previous section.

To facilitate our discussion we shall define a sampling method as consisting of a specific percent p and a specific number of sample clusters m_1, m_2 and m_3 . Because of the limitations due to computer time and costs, we have (i) selected 21 specific sampling methods (in addition to the one analyzed in the previous section) and (ii) eliminated the estimators by Procedures 3, 6, and 7 that required an extremely long computer time and did not seem to present any real advantage over Procedures 2, 4 and 5 respectively. The 22 sampling methods analyzed in this study are listed in Table 7. Note that the

Table 6 - The sample values of the ratio $\sqrt{V/S_{ZZ}}$ arranged by population, regression equation form, procedure number and least squares estimation approach.

Estimation Approach	Form of Regression	Population	Procedure Number								
			1	2	3	4	5	6	7	8	9
OLS	1	1	.71	.68	.71	.68	.67	.72	.71	.68	.67
		2	.73	.68	.73	.68	.67	.73	.72	.73	.72
		3	.79	.68	.72	.68	.67	.72	.71	.72	.71
		4	.77	.68	.72	.68	.67	.73	.72	.73	.71
	2	1	.73	.68	.70	.67	.66	.67	.67	.66	.64
		2	.81	.68	.69	.67	.66	.67	.67	.74	.72
		3	.84	.68	.70	.67	.66	.68	.68	.75	.73
		4	.77	.68	.69	.67	.66	.67	.67	.75	.73
OWLS	1	1	.73	.73	.73	.74	.73	.74	.73	.68	.66
		2	.79	.73	.76	.74	.73	.76	.75	.80	.79
		3	.95	.73	.70	.74	.73	.71	.70	.80	.85
		4	.91	.73	.70	.74	.73	.72	.70	.86	.83
	2	1	.80	.74	.74	.75	.74	.78	.77	.70	.69
		2	.93	.74	.81	.75	.74	.81	.81	.97	.95
		3	1.08	.74	.70	.75	.74	.72	.70	1.07	1.05
		4	.97	.74	.75	.75	.74	.78	.75	.91	.89
MLS	1	1	.97	.86	.93	.87	.85	.93	.92	.92	.90
		2	1.03	.86	.84	.87	.85	.83	.83	1.05	1.02
		3	1.15	.86	.92	.87	.85	.92	.92	1.08	1.06
		4	1.09	.86	.62	.87	.85	.61	.61	1.06	1.03
	2	1	.98	.91	.94	.89	.87	.94	.92	.90	.80
		2	1.11	.91	.89	.89	.87	.85	.85	1.08	1.06
		3	1.31	.91	.93	.89	.87	.91	.91	1.23	1.20
		4	1.18	.91	.89	.89	.87	.86	.86	1.21	1.16
MWLS	1	1	1.09	1.04	1.06	1.05	1.03	1.04	1.06	1.02	.99
		2	1.15	1.04	.99	1.05	1.03	1.01	.98	1.19	1.16
		3	1.39	1.04	1.06	1.05	1.03	1.06	1.06	1.30	1.28
		4	1.35	1.04	.77	1.05	1.03	.79	.78	1.29	1.24
	2	1	1.20	1.12	1.15	1.10	1.10	1.16	1.15	1.07	1.05
		2	1.32	1.12	1.11	1.10	1.10	1.11	1.11	1.34	1.30
		3	1.52	1.12	1.17	1.10	1.10	1.20	1.18	1.54	1.49
		4	1.53	1.12	1.14	1.10	1.10	1.17	1.13	1.53	1.49

sampling method 11 is the method considered in the previous section.

For each population and each sampling method we have calculated the same set of basic statistics as that shown in Table 1. To facilitate the analysis, however, the statistics from the various tables were arranged and rearranged several times in order to answer specific questions or verify specific conclusions reached in the analysis of the previous section. All these tables are listed in Arabatzis (1986).

To better see how the conclusions of the previous section change, we have summarized them all as a set of five main conclusions. Each conclusion will be stated separately and its validity analyzed in view of the new information provided by the results of the 21 additional sampling methods.

Conclusion 1 - (sampling method 11). The estimators by all procedures are unbiased for population 1, the only population for which the differences among strata are negligibly small. There is one exception, however, that of the OWLS estimators of the regression form 2; but their bias is thought to be due to sampling error. For the other three populations, only the procedures 2, 5, and 9 yield unbiased estimators; the procedure 2 that is expected to be unbiased on theoretical considerations, and the two procedure 5 (derived from procedure 2) and 9 that use weights based on stratum total number of trees. Procedures 3 and 7 yield mixed results; the least squares but not the weighted least squares estimators are generally unbiased. Procedure 1 (which ignores the effect of stratification) and procedures 4, 6 and 8 (that are all based on weights defined in terms of stratum areas) yield biased results.

Table 7 - The number of clusters m_1 , m_2 and m_3 of stratum 1, 2 and 3 respectively and the percent p of trees subsampled for the twenty-two stratified sampling methods considered in this study

Sampling Method					Sampling Method				
m_1	m_2	m_3	p	m_1	m_2	m_3	p		
1	5	5	5	15	12	30	30	30	15
2	5	10	20	15	13	5	5	5	5
3	5	15	30	15	14	15	15	15	5
4	10	20	30	15	15	30	20	10	5
5	15	15	15	15	16	30	30	30	5
6	15	20	5	15	17	30	15	5	5
7	15	30	5	15	18	5	5	5	60
8	20	10	5	15	19	15	15	15	60
9	20	30	10	15	20	30	20	10	60
10	30	15	5	15	21	30	30	30	60
11	30	20	10	15	22	30	15	5	60

The analysis of the bias of the entire set of 22 sampling methods shows similar results. We still have the estimators of population 1 unbiased, except for the OWLS estimators by regression form 2. For the other populations, procedures 2 and 5 yield unbiased estimators; but now, some of the results by the procedure 9 are biased. And as before, the bias of the estimators calculated by the procedures 1, 4 and 8 is significantly different from zero with many of the values t being very large.

Conclusion 2 - (sampling method 11). Within a given procedure (i) the weighted least squares estimators are, more often than not, more precise, but not by much, than the least squares ones, (ii) the ordinary least and weighted least squares approaches yield always more precise estimators than the modified ones and (iii) the estimators based on the regression of biomass on diameter and height are always more precise than those based on the regression of biomass on diameter alone.

The analysis of the results from the entire set of 22 sampling methods shows that, with the exception of two special cases that extend them, these conclusions are verified. The two exceptions refer to (i) the case of small number of sample clusters per stratum ($m=5$ results in modified least and weighted least squares regressions having only two degrees of freedom) where the estimators by the modified procedures are much less precise than those by the ordinary procedures and (ii) the case of small number of sample trees per cluster (when $p=5$, the average number of trees per cluster is less than 2) where the precision of the modified least squares method approaches that of the ordinary least squares. This seems to imply that the MLS and MWLS estimators should only be used when there are at least ten sample clusters per stratum.

Conclusion 3 - (sampling method 11). With the possible exception of population 1 (where the strata are very similar), it seems that, although

the estimators by procedure 2 are less precise than those by procedures 1 and 3, they are usually more accurate. Similarly, procedures 5 and 9 may yield slightly better estimators than procedures 2 but they require the knowledge of the total number of trees contained in each stratum. Finally, all of the remaining procedures may be more precise but, because their bias is generally large and, sometimes, they are inconsistent, their accuracy is low.

These conclusions are all verified by the analysis of the results from the additional 21 sampling methods.

Conclusion 4 - (sampling method 11). The sample-based statistic V grossly underestimates the variance of z when the ordinary least or weighted least squares techniques are used; the underestimation is much smaller with the modified least squares approach and for the modified weighted least squares, V overestimates the variance of z by a relatively small amount.

The analysis of the other 21 additional sampling methods show that, with some exceptions that may be viewed more as extensions, these conclusions hold true. When the number of sample trees per cluster is small (the case of $p=5$ which results in an average of less than 2 trees per cluster) the underestimation of the variance of z becomes smaller, and for the case of OWLS estimators it becomes negligibly small. When the number of sample trees per cluster increases to about 20 (the case of $p=60$) the underestimation of the precision of the OLS and OWLS estimators becomes much larger; often, the standard deviation of z is estimated to less than 50 percent of its value. As far as the modified least and weighted least squares methods are concerned, there seems to be no effect of p on the underestimation or overestimation of the precision.

Conclusion 5 - (sampling method 11). Taking into consideration the bias, precision and estimated precision, it seems that the best overall estimation procedure is procedure 2; it is unbiased, generally less precise but more accurate than procedures 5 and 9 which, although they may yield about equally good results, would require the additional knowledge of stratum size expressed as total number of trees. This conclusion holds true for the additional 21 sampling methods as well.

Summary Comments

We have described a specific stratified, two-stage sampling method of tree selection for construction of tree biomass tables for use with forest resource inventory. This method consists of (i) a first stage sample of m_1 , m_2 and m_3 clusters (plots) selected from stratum 1, 2 and 3 respectively of a given forest population and (ii) a second stage subsample of p percent of the trees contained in each sample cluster of the first stage. The selection procedure in both stages is simple random sampling without replacement. To facilitate the discussion, we have

defined a sampling method as consisting of a specific sample size and allocation, that is, a specific set of sampling parameters m_1 , m_2 , m_3 and p .

Using simulation techniques, these sampling methods were then repeatedly applied to a given forest population of 22,723 trees distributed in 927 one-fifth acre sample plots. These plots were part of the New York State forest inventory system carried out by the Northeastern Forest Experiment Station. The trees were measured in the field for their diameter and merchantable height, but not measured for their total height and biomass (green weight above ground); these were generated by a Monte Carlo process that preserved the effect (on the simulated value of the tree height and biomass) of such factors as species, diameter, merchantable height, site, geographical region, plot and inherent random variation. Although artificially constructed, this population of trees is expected to imitate, with sufficient accuracy, a natural forest; that is, is expected to have all of the basic characteristics of forest populations as found in nature. The forest population so constructed was then divided, somewhat arbitrarily into three geographical regions called strata. As the biomass regressions of the three resulting strata were quite similar, we have constructed three additional forest populations with biomass regression functions that were made to vary greatly from stratum to stratum.

The computer simulation process of sampling was repeated one hundred times and, thus, for each sampling method we have obtained one hundred different samples of trees. The tree biomass regression function of each sample was then estimated with the use of seventy-two statistical models, the combinations of four estimation approaches (OLS, OWLS, MLS and MWLS), two regression function forms (1 and 2) and nine procedures that took the stratum effect into account; procedure 1 that ignores the stratum effect and calculates one regression function for all strata, procedure 2 that calculates independent regressions for each stratum separately, procedure 3, similar to procedure 2, that uses dummy variables techniques to calculate individual stratum regressions that are not statistically independent of each other and the remaining six procedures that calculate single regressions for use in all strata by either (i) appropriately averaging the individual regressions of procedures 2 and 3 or (ii) appropriately weighing the information from sample tree data when the weighted least squares method is being applied.

The seventy-two regression functions of each sample of a representative set of sampling methods were then applied to the tree data of the four populations to calculate estimators z of (the average biomass per acre) and estimators S_{zz} and V of the variance of z . The estimator S_{zz} is calculated from the one hundred, identically distributed values z of the same sampling and estimation procedure, and the estimator V is calculated from the data of a single sample, under the assumptions of the statistical model

used. V is the only estimator of the variance of z that one can obtain in real life, but this estimator is generally biased as the assumptions of the model are rarely satisfied. On the other hand, S_{zz} is an unbiased estimator of the variance of z but it is seldom, if ever, known in real life. Of course, we can calculate S_{zz} by deliberately sampling the same population several times by the same sampling method, or we can deliberately split the sample data into several parts and use the variation between the estimates z of the parts as the basis for the calculation of S_{zz} .

Because the true value μ is known and we have one hundred, statistically independent and identically distributed sets of random variables z , V and $(z+t\sqrt{V})$, we are able to study the probability behavior of these variables for any given combination of sampling method and estimation procedure. In particular we can estimate the bias of z by the quantity $(z-\mu)$ and the accuracy of z by the mean square error defined as the sum of the variance S_{zz} and squared bias $(\bar{z}-\mu)^2$. We can also test the null hypothesis that the bias of z is equal to zero, see whether V is a valid estimator of the variance of z , or whether the confidence intervals based on V are valid statistical inferences. Finally, the variance S_{zz} , the corresponding mean square $(S_{zz}+(\bar{z}-\mu)^2)$ or their square roots can be used to compare the efficiency of various combinations of sampling methods and estimation procedures.

The analysis of the simulation results led to the several important conclusions listed in the previous section. In particular one should note that (i) the ordinary least squares method is almost as good as the ordinary weighted least squares but they both underestimate the error of the estimators; the only exception being the case of $p=5$ (less than two trees, on the average per cluster) where the error of the OWLS (but not that of the OLS) estimators is properly evaluated by V , (ii) the modified least and weighted least squares estimators are somewhat less precise than the corresponding ordinary ones, when the number of sample clusters within a given stratum is sufficiently large, say greater than ten; the modified estimation approaches should never be used, however, when the number of sample clusters is small, say less than ten, (iii) the usual procedure of calculating the biomass regression function by combining the data of the trees from all strata (procedure 1) is dangerous; it may become an important source of bias, (iv) the best overall procedure in terms of bias and accuracy is that of constructing biomass regressions separately by stratum (procedure 2) and then applying the usual stratified sampling formulae to derive estimators and their error.

The surprising conclusion about the equal efficiency of least and weighted least squares method should be properly interpreted. They are equally good in terms of the bias and the precision of the estimator of the average biomass per unit area. If one wishes to calculate a biomass regression function with valid estimates of its precision (confidence and prediction in-

tervals) one must use the weighted least squares method.

Acknowledgements

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

Literature Cited

- Arabatzis, A. A. Evaluating the error of tree biomass regressions by simulation; sample trees selected by stratified cluster sampling. M.S. thesis, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Cunia, T. Dummy variables and some of their uses in regression analysis. In: Proceedings of the June 1973 Meeting of IUFRO Subject Group S4.02, Nancy-France, Vol. 1, T. Cunia, K. Kuusela, and A. J. Nash (Eds.), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1973.
- Cunia, T. Evaluating errors of tree biomass regressions by simulation. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Cunia, T.; Gillespie, A. J. Cluster sampling and construction of biomass tables: results of a simulation study. In: Proceedings, third annual southern forest biomass workshop, March 12-14, 1985; University of Florida, Gainesville, FL; 1985.
- Cunia, T.; Michelakackis, J. A method to construct a forest biomass population model. In: Proceedings, Renewable resource inventories for monitoring changes and trends. J. F. Bell and T. Atterbury (Eds.), Oregon State University, Corvallis, OR; 1983.
- Cunia, T.; Michelakackis, J. A Monte Carlo technique for generating total height of forest trees. Faculty of Forestry Miscellaneous Publication Number 4 (ESF 84-018), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1984a.
- Cunia, T.; Michelakackis, J. Constructing forest biomass populations for simulated sampling. Faculty of Forestry Miscellaneous Publication Number 5 (ESF 84-019), SUNY College of Environmental Science and Forestry, Syracuse, NY; 1984b.
- Cunia, T.; Michelakackis, J.; Lee, S. Generating total tree heights by a Monte Carlo technique. In: Proceedings, 1983 southern forest biomass workshop, June 15-17, 1983, Charleston, SC, R. F. Daniels and P. H. Dunham (Eds.), USDA Forest Service, Southeastern Forest Experiment Station, Asheville, NC; 1984.
- Gillespie, A. J. Estimation of biomass tables by cluster sampling: a simulation study. M.S. thesis, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1985.
- Gillespie, A. J.; Cunia, T. Error of biomass regressions: sample trees selected by cluster sampling. In: Proceedings of the Workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates". May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.
- Michelakackis, J.; Cunia, T. Construction of biomass tables by double sampling: preliminary results of a simulation study. In: Proceedings, Use of auxiliary information in natural resource inventories, October 1-2, 1985, Blacksburg, VA. R. G. Oderwald, H. E. Burkhardt and T. E. Burk (Eds.), Society of American Foresters Publication No. SAF 86-01; 1985.
- Michelakackis, J.; Cunia, T. Error of biomass regressions: sample trees selected by double sampling. In: Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.