Sampling forest plots using adaptive empirical cumulative distribution functions

John Tipton

January 28, 2014

1 Introduction

In the study of paleoclimate and paleoecology, tree ring chronologies are an invaluable proxy. For the construction of stand level tree ring chronologies, the selection of which trees within a plot to take tree core samples of is vital. Methods for selecting trees for coring include selecting the ten largest trees on the plot, sampling line quadrats, others?? **Citations and References - Neil, Ross, etc.** Many of the methods are either non-probabilistic and therefore lack desirable statistical properties or are probabilistic but don't adequately capture the larger trees in the skewed tail of the distribution of tree size. This drives the interest is the development of a flexible, adaptable method that can be employed in the field under the constraints of a probabilistic sampling protocol.

Due to in field constraints, the sampling mechanism will not know the ancillary variable on which to inform the sample, but this value will be measured as the sampling scheme progresses. The aim of this manuscript is to outline a sampling scheme that is easy to implement in the field but gives desired sampling properties, namely the design biases sampling to include larger trees while still maintaining statistical properties of unbiasedness for estimating the total plot biomass.

2 Distribution and description of data

To understand the goals of the proposed sampling design, a description of the data is needed. DBH (diameter at breast height) is a measure of the diameter of a tree at 1.4 m above the ground and is treated as the ancillary variable on which to sample. For a given species, the distribution of DBH is strongly right skewed. This is important as the trees most valuable for paleoclimate and paleoecology reconstructions are those with large dbh values. These large dbh value trees are important as they are likely older (and therefore having more rings and a longer chronology) and also have more biomass (and therefore play a greater role in carbon sequestration at the stand level). Figure 1 shows a simulated distribution of dbh that is reflective of the skewness in forest plots.

DBH is used as a predictor for biomass using an allometric equation. Under simple random sampling, the model of interest is the allometric tree growth equation with exponential error given by

$$Y_i = X_i^{\beta_1} e^{\beta_0 + \epsilon} \tag{1}$$

where Y_i is biomass for tree i, X_i is dbh for tree i, $\epsilon \sim N(0, \sigma^2)$ is a random error, and β_0 and β_1 are coefficients to be estimated. A plot of the allometric relationship using simulated data is seen in Figure 2.

The allometric model is estimated by using ordinary least squares on the log-log model

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + \epsilon \tag{2}$$

Figure 1: Plot of simulated DBH typical of a sample plot (assumes one species only)

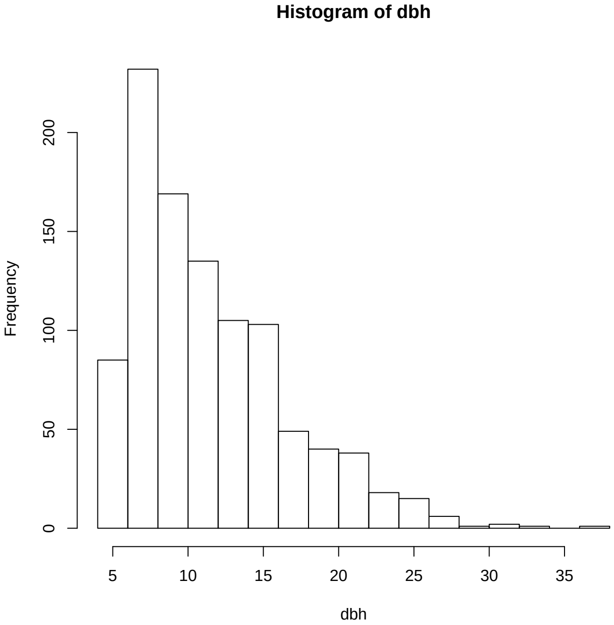


Figure 2: Plot of simulated allometric relationship typical of a sample plot (assumes one species only)

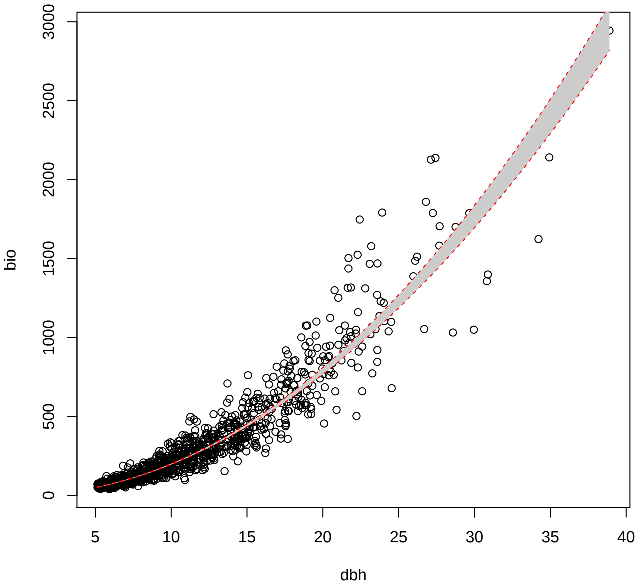
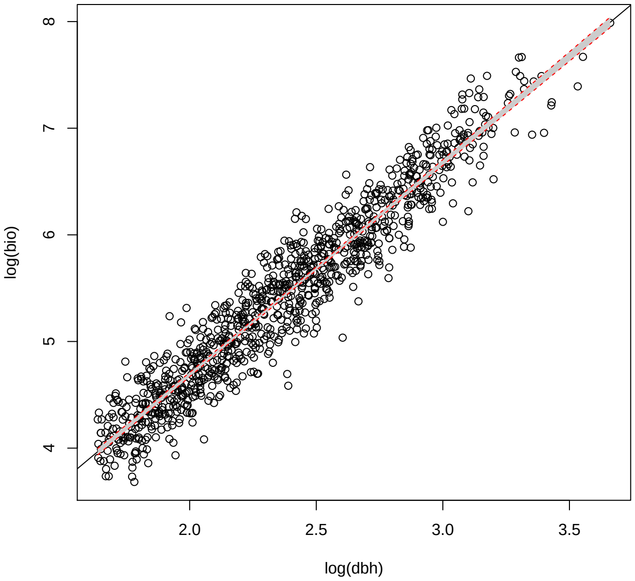


Figure 3: Plot of simulated log-allometric relationship typical of a sample plot (assumes one species only)



A plot of the log-log model is shown in Figure 3. After fitting model (2), the biomass is then back-transformed into the exponential model (1) while accounting for the variance with the use of a Delta method like transformation.

There are three primary questions to be addressed: one, what is the effect of stratified sampling on the estimate of total biomass and its variance estimator, two, what is the effect of stratified sampling on the estimation of the coefficients β_0 and β_1, and the estimation of the variances associated with the coefficients β_0 and β_1, and three, can the stratification scheme be optimized for the desired sampling goals of estimating the total plot biomass with as small variance as possible while biasing sampling towards larger trees. The answer to the first question about estimating total biomass and its corresponding variance estimator will require a regression estimator and knowledge of the design based sampling weights $\boldsymbol{\pi}$. The weights $\boldsymbol{\pi}$ will be discussed later in the manuscript. The Second question about estimation of the parameter vector $\boldsymbol{\beta}$ is important as the uncertainty interval for total biomass for a given DBH is important in further ecological modeling processes. Of interest is the "fanning out" of the confidence interval for biomass as DBH increases as seen in Figure 2. The third question can be discussed in detail after unbiasedness of the first two conditions is established.

3 Finite Population Design

Consider a finite population of size N derived from a continuous distribution. Denote this population $U = \{1, \ldots, N\}$. A simple and easy way to reduce variance in the estimation of a population total is the use of stratified sampling. Stratified sampling is accomplished by sampling from groupings of the ancillary variable X known a priori. For instance, if the values of X are grouped small, medium, and large, a stratified design consists of sampling n_1 elements form the small group, n_2 elements from the medium group, and n_3 elements from the large group. In general, this is extended to h strata where the size of stratum h is known and denoted N_h and the population size is

$$N = \sum_{h=1}^{H} N_h.$$

The population total is

$$\begin{aligned} t_y &= \sum_{i \in U} y_i \\ &= \sum_{h=1}^{H} t_h \\ &= \sum_{h=1}^{H} N_h \bar{y}_h \end{aligned}$$

where t_h is the stratum total and $\bar{y}_h$ is the stratum mean. In stratified sampling, the π estimator for the population total is

$$\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi}$$

where $\hat{t}_h$ is the π estimator of t_h. The variance of the estimator $\hat{t}_\pi$ is given by

$$V(\hat{t}_\pi) = \sum_{h=1}^{H} V(\hat{t}_{h\pi})$$

where $V(\hat{t}_{h\pi})$ is the stratum variance of the stratum total $\hat{t}_{h\pi}$. An unbiased variance estimator for the stratified design is

$$\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} \hat{V}(\hat{t}_{h\pi})$$

where $V(\hat{t}_{h\pi})$ is the estimated stratum variance. Within each stratum h, a simple random sample of size n_h can be taken. Under this design s_h for sampling stratum h, the π estimator for the population total $\sum_{i \in U} y_i$ is

$$\hat{t}_\pi = \sum_{h=1}^{H} N_h \bar{\boldsymbol{y}}_{s_h}$$

where $\bar{\boldsymbol{y}}_{s_h} = \sum_{i \in s_h} \frac{y_i}{n_h}$ is the sample mean for stratum h. The variance for the estimator of the population total is

$$V(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} s^2_{yU_h}$$

where $f_h = \frac{n_h}{N_h}$ is the sampling fraction in stratum h and

$$s^2_{yU_h} = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_i - \bar{y}_{U_h})^2$$

is the stratum variance and $\bar{y}_{U_h} = \sum i \in U_h \frac{y_i}{N_h}$ is the stratum mean. An unbiased estimator of the variance is

$$\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} s^2_{ys_h}$$

where

$$s^2_{ys_h} = \frac{1}{n_h - 1} \sum_{i \in s_h} (y_i - \bar{y}_{s_h})^2$$

is the stratum variance in stratum h. The mean plot biomass can then be calcualted as $\frac{\hat{t}_\pi}{N}$ and the variance estimator for the mean plot biomass is $\frac{1}{N^2}\hat{V}(\hat{t}_\pi)$.

4 Sampling Design

1. Count (roughly estimate) the number of trees on the plot of interest in each stratum category of interest (e.g. small, medium large).

2. Based on the goals of the study, divide the sampling effort n into the categories of interest, making sure to keep a minimum of 5? elements per stratum category.

The model (2) could be estimated by using weighted least squares regression where the weight matrix $W = diag(\hat{\pi}_1, \ldots, \hat{\pi}_N)$ and this could then be transformed into model (1) by appropriately transforming the variance estimates through a Delta method like transform.

5 Simulation and results

For the simulation into investigating the sampling schemes, five different sampling schemes were considered, simple random sampling (SRS), probability proportional to size sampling (PPS), sampling using the empirical cumulative distribution function (ECDF), sampling using an adaptive estimation of the empirical cumulative distribution function (AECDF), and stratified sampling with simple random sampling within each stratum (STSI). Assuming that the goals of the sampling scheme are one: increase the number of large dbh trees in the sample to provide a longer historical record for climate reconstruction in a statistically valid way, two: reduce the uncertainty in estimating total plot biomass from the allometric relationship between dbh and biomass relative SRS sampling, three: have a simple sampling protocol that can be implemented in the field, the five sampling schemes can be compared.

SRS sampling is the best known and one of the easiest sampling schemes to implement in the field as it only requires knowing (or estimating) the number of trees on the plot, but it fails to preferentially select large dbh trees and it is the basis for comparison for a reduction of variance so will not be efficient.

PPS sampling assigns sample inclusion probabilities based on a size variable known a priori, in our case we would use dbh if it was known. PPS sampling assigns sample weights for element i of $\pi_i = \frac{x_i}{\sum_{i \in U} x_i}$. This presents problems for implementation in the field as it requires measuring all of the dbh values on the plot and labeling the trees. This labeling could be done in practice, but might be prohibitive in cost of time but will be highly efficient, in fact this scheme will reduce variance estimates more than any other method under the proposed model (1). PPS sampling will also preferentially sample larger dbh trees relative to smaller dbh trees. PPS sampling meets criteria one and two but not three

ECDF sampling assigns sample inclusion probabilities based on the finite population empirical distribution function $F(\cdot)$ where $F(x_{(n)}) = \frac{n}{N}$ for $x_{(n)}$ the n^{th} order statistic. Like PPS sampling, this requires knowing the values of the dbh a priori, which is not possible in this particular example. The ECDF sampling design does provide a point of comparison of the AECDF method as the ECDF is a best case scenario of the AECDF method (i.e. if the ECDF method is not sufficient for this problem, then the AECDF design is necessarily not sufficient). The ECDF method will reduce the variance relative to SRS and will preferentially sample larger trees, thus meeting criteria one and two but not three.

AECDF sampling assigns sample inclusion probabilities adaptively so that in the superpopultion model asymptotics the design is equivalent to ECDF sampling. AECDF sampling has the advantage in that it can be easily implemented in the field using modern technology like a tablet computer and excel spreadsheet. At each tree dbh measurement, a up/down sampling decision can be made and in the superpopulation model these sampling decisions approximate ECDF sampling without knowing a priori the values of dbh. AECDF also preferentially samples larger dbh trees. For finite populations, the question of interest is the effect of the AECDF design on the estimation of plot level biomass. The AECDF design meets criteria one and three, and the effect of this design on criteria two is of interest.

Table 1: Bias and Variance for design based estimates of mean plot biomass

	Bias	Relative Efficiency
SRS	-0.00	1.00
ECDF	-0.02	0.67
PPS	-0.10	0.12
AECDF	0.00	1.35
STSI	-0.00	0.59

STSI sampling assigns sample inclusion probabilities for different size classes according to a simple random sampling scheme. Stratified sampling has the advantage of grouping similar elements of the population and thus reducing the variance of estimates by reducing within group variances. STSI sampling can also be made to preferentially sample larger trees by preferentially sampling the largest dbh class. STSI sampling can easily be implemented in field by initially estimating the number of trees in each size class on the plot and then using a random number generator in a tablet PC or smartphone to randomly select the sample within each size class. The STSI meets all three of the design criteria of interest and its performance relative to other designs is of interest.

Using simulated data for dbh and biomass as shown in Figures 1-3 as a finite population approximating the true population criteria two can be evaluated for the different designs. From Table 1, it is seen that the most efficient design is PPS sampling with a relative efficiency of 0.12 for this particular simulation. The ECDF and STSI designs perform relatively similarily with respect to the relative efficiency measure with the STSI design being easier to implement without prior knowledge of the distribution of dbh. The AECDF method has relative efficiency higher than SRS and would only be useful if design criterions one and three are of interest.

6 Appendix - Current discussion items, thoughts, hypotheses, and other trains of thought

- First, we need to pin down the primary questions of interest and define the goals
  - Is the goal to sample large trees that will likely have a longer chronology and thus allow a better back forecast of biomass in a probabilistic way?
  - Is the goal to decrease the uncertainty about biomass for large values of dbh?
  - Is the goal to develop an adaptive sampling design that can be implemented in field to preferentially select larger trees in a statistically sound way while maintaining a probabilistic sample or is it to design a robust sampling design that is easy to implement?
- Does the design need to incorporate multiple species? perhaps you could stratify based on size and species?