Sampling forest plots using adaptive empirical cumulative distribution functions

John Tipton

February 4, 2014

Introduction

Goals

- Estimate plot level biomass with increased precision
- Increase the number of large dbh trees in the sample
- Easily implementable in the field

Biomass Model

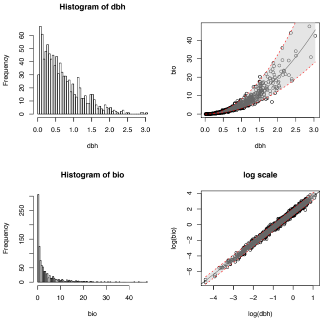
DBH is used as a predictor for biomass using an allometric equation. Under simple random sampling, the model of interest is the allometric tree growth equation with exponential error given by

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

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.

The Model

A plot of the allometric relationship using simulated data is seen below

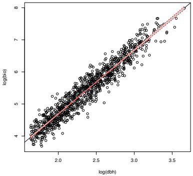


The Model

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

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + \epsilon \quad (2)$$

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



Three questions:

1. What is the effect of sampling design on the estimate of total biomass and its variance estimator

2. What is the effect of sampling desing 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

3. How to optimize the sampling design for the desired sampling goals of stimating the total plot biomass with high precision while selecting large trees

Horvitz-Thompson Estimation

For a finite population with elements indexed by $i = 1, \ldots, N$ with a probability π_i of being included in the sample s, the unbiased estimate of the population total is

$$\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}$$

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 log model can be estimated by using weighted least squares regression with the weight matrix $W = diag(\hat{\pi}_1, \ldots, \hat{\pi}_N)$
- The estimates then be transformed into by appropriately transforming the variance estimates through a Delta method like transformation.

Six different sampling schemes were considered

1. Simple random sampling (SRS)
2. Probability proportional to size sampling (PPS)
3. Empirical cumulative distribution function (ECDF)
4. Adaptive estimation of the PPS design (APPS)
5. Adaptive estimation of the ECDF (AECDF)
6. Stratified sampling with simple random sampling within each stratum (STSI)

Comapring the designs 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

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.

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 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 similarly 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.

Table : 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

- 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?