Sampling forest plots using adaptive empirical cumulative distribution functions

John Tipton

February 5, 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

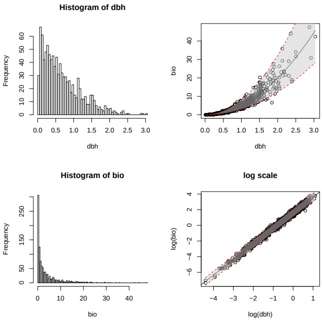
DBH is used as a predictor for biomass using an allometric equation. Under simple random sampling, the model of interest is the allometric tree growth equation with exponential error given by

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

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



The Model

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

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

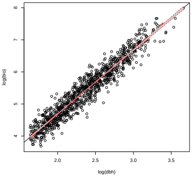


Figure : Plot of simulated log-allometric relationship typical of a sample plot (assumes one species only)

Questions of Interest

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

$$\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

Estimation of the variance of the population total is calculated as

$$\hat{V}(\hat{t}_\pi) = \sum_{k \in s} \sum_{l \in s} \frac{1}{\pi_{kl}} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) y_k y_l$$

- The log model can be estimated by using weighted least squares regression with the weight matrix $W = diag(\hat{\pi}_1, \ldots, \hat{\pi}_N)$
- Weighted least squares regression model estimate for β is $\hat{\beta} = (X^T WX)^{-1} X^T Wy$
- 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. Empirical cumulative distribution function (ECDF)
3. Probability proportional to size sampling (PPS)
4. Adaptive estimation of the ECDF (AECDF)
5. Adaptive estimation of the PPS design (APPS)
6. Stratified sampling with simple random sampling within each stratum (STSI)

Simple random sampling (SRS)

- Easy to implement in the field
- Fails to preferentially select large dbh trees
- Baseline for comparison of precision

Probability Proportional to Size (PPS)

- PPS sampling assigns sample inclusion probabilities based on a size variable known a priori, in our case dbh
- PPS sampling assigns sample weights for element i of $\pi_i = \frac{dbh_i}{\sum_{i \in U} dbh_i}$
- Requires measuring all of the dbh values on the plot and labeling the trees before coring
- Highly efficient. This scheme will increase precision more than any other method
- PPS sampling will also preferentially sample larger dbh trees relative to smaller dbh trees

Empirical Cumulative Distribution Function Sampling ECDF

- 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
- This requires knowing the values of the dbh a priori
- ECDF provides a point of comparison of the AECDF method as the ECDF is a best case scenario of the AECDF method
- ECDF will reduce the variance relative to SRS
- ECDF will preferentially sample larger trees

Stratified Sampling (STSI)

- STSI sampling assigns sample inclusion probabilities for different size classes according to a simple random sampling scheme
- STSI groups similar elements of the population and thus increases the precision of estimates by reducing within group variances
- STSI sampling can be made to preferentially sample larger trees
- 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

Adaptive ECDF sampling (AECDF)

- AECDF asymptotically approaches the ECDF design
- AECDF sampling can be implemented in the field using a tablet or smartphone
- At each dbh measurement, an up/down sampling decision to core can be made without knowing all the values of dbh
- AECDF 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
- AECDF variance estimator will need to be derived mathematically (not currently solved to the best of my knowledge)

Adaptive PPS sampling (APPS)

- APPS asymptotically approaches the PPS design
- APPS sampling can be implemented in the field using a tablet computer or cell phone
- At each dbh measurement, an up/down sampling decision to core can be made without knowing all the values of dbh
- APPS preferentially samples larger dbh trees
- For finite populations, the question of interest is the effect of the APPS design on the estimation of plot level biomass
- APPS variance estimator will need to be derived mathematically (not currently solved to the best of my knowledge)

Using the simulated data presented earlier, I repeatedly sampled $n = 10, 40$, and 100 trees from the simulated population of $N = 400$ trees using the proposed designs and calculated Monte Carlo bias and relative efficiencies for each design based on 10,000 replicates.

Results $n = 20$

	Bias	Relative Efficiency
SRS	-0.00	1.00
ECDF	-0.01	0.71
PPS	-0.04	0.13
AECDF	0.06	1.40
APPS	-0.02	0.47
STSI	0.00	0.48

Results $n = 40$

	Bias	Relative Efficiency
SRS	-0.02	1.00
ECDF	-0.00	0.73
PPS	-0.07	0.14
AECDF	0.01	1.41
APPS	0.00	0.36
STSI	-0.00	0.37

Results $n = 100$

	Bias	Relative Efficiency
SRS	0.01	1.00
ECDF	0.00	0.54
PPS	-0.27	0.10
AECDF	0.04	1.45
APPS	-0.00	0.56
STSI	0.00	0.55

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.

- 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?