Remote Sensing of Environment 197 (2017) 85–97

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Hybrid three-phase estimators for large-area forest inventory using ground plots, airborne lidar, and space lidar

Sören Holm [a,1], Ross Nelson [b,*,1], Göran Ståhl [a]

[a] Dept. of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå SE-90183, Sweden
[b] Code 618, Biospheric Sciences Laboratory, NASA - Goddard Space Flight Center, Greenbelt, MD 20771, USA

ARTICLE INFO

Article history:
Received 27 September 2016
Received in revised form 29 March 2017
Accepted 11 April 2017
Available online xxxx

Keywords:
Hybrid 3-phase estimation
Large-area forest inventory
Biomass estimation
Satellite lidar

ABSTRACT

Previous studies have utilized ground plots, airborne lidar scanning or profiling data, and space lidar profiling data to estimate biomass across large regions, but these studies have failed to take into account the variance components associated with multiple models because the proper variance equations were not available. Previous large-domain studies estimated the variances of their biomass density estimates as the sum of the GLAS sampling variability plus the model variability associated with the models that predict airborne lidar estimates of biomass density (Y) as a function of satellite lidar measurements (X). This approach ignores the additional variability associated with the predictive models used to estimate ground biomass density as a function of airborne lidar measurements. This paper addresses that shortcoming. Analytic variance expressions are provided that include sampling variability and model variability in situations where multiple models are employed to generate estimates of biomass. As an example, the forest biomass of the continental US is estimated, by forest stratum within state, using a space lidar system (ICESat/GLAS). An airborne laser system (ALS) is used as an intermediary to tie the GLAS measurements of forest height to a small subset of US Forest Service (USFS) ground plots by flying the ALS over the ground plots and, independently, over individual GLAS footprints. Two sets of models are employed to relate satellite measurements to the ground plots. The first set of equations relates USFS ground plot estimates of total aboveground dry biomass density (Y_1) to spatially coincident ALS forest canopy measurements (X_1). The second set of models predicts those ALS canopy height measurements (X_1) used in the first set of models to GLAS waveform measurements (X_2). The following important conclusions are noted. (1) The variability associated with estimation of the plot-ALS model coefficients is significant and should be included in the overall estimate of biomass density variance. In the continental US, the total variance of mean forest biomass density (98.06 t/ha) increases by a factor of 3.6×, i.e., from 1.91 to 6.94 t^2/ha^2, when plot-ALS model variance is included in the calculation of total variance. (2) State-level results are more variable, but on average, the percent model variance at the state level, i.e., (model variance / total variance) ∗ 100, increases from 16% to 59% when plot-ALS model variance is included. (3) The overall model variance is driven in large part by the number of plots overflown by the ALS and the number of GLAS pulses overflown by the ALS. Given a choice of improving precision by either increasing the number of plot-ALS observations or increasing ALS-GLAS observations, there is no obvious benefit to selecting one over the other. However, typically the number of ground plots overflown is the limiting factor. (4) If heteroskedasticity is evident in either the ground-air or air-satellite models, it can modeled using weighted regression techniques and incorporated into these model variance formulas in straightforward fashion. The results are unambiguous; in a hybrid three-phase sampling framework, both the ground-air and air-satellite model variance components are significant and should be taken into account.

1. Introduction

Decades of airborne and space lidar studies have indicated that lidars can be used to estimate forest biomass and carbon, and experience has shown that large-area estimates of biomass density means and biomass totals are easily calculated. The challenge lies with the derivation of the associated variance estimators. As noted by Gregoire et al. (2016), too many of the proposed variance "solutions" have been ad-hoc, utilizing variance estimators not firmly founded on statistical principles or that have accounted for only a portion of the actual variance of the estimates. We believe that the lack of statistical rigor discussed by Gregoire et al. is due in large part to the fact that technology has run ahead of the remote

* Corresponding author.
E-mail address: rfn104@gmail.com (R. Nelson).
[1] Retired.

http://dx.doi.org/10.1016/j.rse.2017.04.004
0034-4257/Published by Elsevier Inc.

sensing community's collective ability to develop robust, statistically defensible estimators that can handle the remarkable airborne and space technology available to us.

For example, Neigh et al. (2013) and Margolis et al. (2015) used a model-based, three-phase sampling approach to inventory forest biomass and carbon in the circumpolar boreal forests and the North American boreal forests, respectively. In order to use ICESat/GLAS (Ice, Cloud, and land Elevation Satellite/Geosciences Laser Altimeter System, https://nsidc.org/data/icesat) to estimate biomass, they had to tie the satellite measurements of canopy height to ground measurements of forest biomass in the far north. These are remote areas with little or no access and thus relatively few ground plots. They used an airborne laser system as an intermediate measurement tool to tie GLAS to ground. The airborne laser acquired ranging measurements on existing ground plots in order to develop models relating ground biomass (Y_{ground}) to airborne lidar (ALS) measurements of canopy height and density (X_{ALS}). The same airborne lidar was flown along 1000 s of kilometers of GLAS orbits, transiting 10's of thousands of individual GLAS pulses. They used the ground-air models to estimate forest biomass on each of the GLAS pulses overflown, thus facilitating derivation of a second set of models which predict airborne laser estimates of forest biomass ($\widehat{Y}_{ALS}$) as a function of GLAS forest canopy height measurements (X_{GLAS}). Once calculated, this second set of equations allows GLAS to be used as a continental sampling tool to estimate forest biomass and carbon across millions of square kilometers of Alaskan, Canadian, Scandinavian, and Russian boreal forest.

Neigh et al. (2013) and Margolis et al. (2015) describe a hybrid 3-phase sampling framework (Ståhl et al., 2016) where the first phase (GLAS) is design-based and the 2nd (ALS) and 3rd (ground plots) phases utilize model-based inference. They employ variance formulas (Ståhl et al., 2011, validated by Ene et al., 2012) that ignore the variability associated with the ground-air models. They did so because the appropriate model-based, three-phase estimator was not available at the time their studies were completed. As they clearly state in their reports, this omission may produce variance underestimates. The primary objective of this study is to rectify this situation, i.e., to provide analysts with analytic variance formulas that account for three error sources: (1) GLAS sampling error, (2) ground biomass-airborne lidar model error, and (3) airborne lidar-GLAS model error. The secondary objective is to empirically demonstrate, via a hybrid three-phase inventory of the forest resources of the US circa 2005, the fact that exclusion of the ground-air error source leads to significant underestimation of the overall variance of stratum, state, and national estimates.

Saarela et al. (2016) report on an approach that may be used to account for multiphase model variances in a situation where the space remote sensing platform (in their case, Landsat ETM+) provides wall-to-wall coverage, i.e., a census of the area of interest (AOI). The current study addresses the situation where the space platform, ICESat/GLAS, samples the Earth's surface, thereby necessitating the need to incorporate both sampling variance and multiple model variance components.

2. Background

Numerous large-area studies have utilized GLAS data as a sampling tool to derive large-area estimates of biomass density and total biomass. Boudreau et al. (2008) attempted to account for sampling variability and for the covariance between GLAS orbital estimates of biomass but ignored the model variance component. Nelson et al. (2009) attempted to account for sampling variability and model variability, but the model variance characterized was that associated with the spread of predicted values about a given regression line, i.e., the regression mean squared error (MSE), not the more important variance component associated with the uncertainty of the model coefficients. Employing the Ståhl et al. (2011) estimators, Neigh et al. (2013), Margolis et al. (2015), and Nelson et al. (2017) improved things by accounting for both GLAS sampling error and ALS-GLAS model variability but had to ignore the variability associated with the plot-ALS models. At the time these studies were completed it was unclear as to how to incorporate that 2nd model (ground plot-ALS) error, hence their AGB model variances are approximate and likely underestimated.

These last three studies in particular provide the impetus to develop hybrid three-phase estimators. In the absence of previously established theory about the sampling design, we denote it as three-phase sampling. However, it does not follow the textbook template of a three-phase design. A textbook three-phase sample design involves collection of a large first-phase sample, e.g., GLAS pulses, and this first phase sample contains auxiliary information (e.g., GLAS height metrics). The second phase subsamples the first, collecting additional auxiliary data, e.g., ALS height and canopy density metrics, on selected 1st phase observations. The third phase, e.g., ground plots, would then be established on a subsample of the second phase sample (e.g. Gregoire and Valentine, 2008) and the variable of interest, e.g., aboveground forest biomass, measured or estimated on each ground plot. Regression or ratio estimation can then be used to tie the three phases together to facilitate large-area estimation. In our case the first phase was the sample of GLAS pulses and the second phase was the subsample of GLAS pulses overflown with the airborne laser scanner. Thus, the first two phases follow the three-phase sampling template. However, our third phase was not a subsample of the second phase sample, but rather an independent sample of field plots over which airborne laser scanner data were collected. To distinguish our design from the standard three-phase case we suggest that it could be denoted three-phase sampling with an independent third phase. But in order to keep the text short, the design is simply denoted three-phase sampling throughout.

A second driver that provides an impetus to develop hybrid three-phase estimators is that there are extensive areas of the Earth's surface that are remote and that host relatively sparse networks of ground observations. These ground plots, often clusters of research or industrial plots rather than probabilistically-designed forest inventories, are needed to calibrate the remote sensing observations to enable prediction of biomass. Such plots can be utilized in a model-based environment but cannot support a design-based study. A recent proprietary study estimates that approximately 60–70% of the Earth's terrestrial surface has no probability-based forest inventory information available on it. These areas include the circumpolar boreal forests with the exception of Scandinavia, tropical and northern Africa, the Middle East, India, and much of SE Asia. More importantly, over the next decade, near-future space lidars such as ICESat-2 (http://icesat.gsfc.nasa.gov/icesat2/) and GEDI (Global Ecosystem Dynamics Investigation, http://science.nasa.gov/missions/gedi/) will be multi-beam profilers, i.e., landscape sampling tools, not imagers. Forest biomass estimates derived using these satellite lidar measurements will have to account for both sampling and model variances.

Ståhl et al. (2011) provide variance estimators that may be used in a hybrid two-phase sampling framework where (1) an airborne or space remote sensing instrument, e.g., an airborne or space lidar, is used to sample (not census) an extensive area-of-interest (AOI), e.g., a province, state, nation, continent, and (2) a model is used to estimate a ground feature of interest, e.g., forest biomass, based on the measurements acquired by the airborne or space lidar. In that work, Ståhl et al. indicate that the total variance of a given regional estimate of biomass is the sum of (A) the sampling variability associated with the lidar acquisition and (B) the model variability that characterizes the effect of the uncertainty in the estimates of the coefficients of the model that predicts ground biomass as a function of the lidar measurements. Their work assumes that (1) the ground measurements are made without error, (2) that the remote sensing measurements constitute a random sample of the landscape, (3) the model form is correctly specified and parameterized, and (4) that the lidar sample acquired to characterize the AOI is independent of the spatially coincident plot-lidar observations used to