Remote Sensing of Environment 192 (2017) 1–11

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A new prediction-based variance estimator for two-stage model-assisted surveys of forest resources

Svetlana Saarela[a,*], Hans-Erik Andersen[b], Anton Grafström[a], Sebastian Schnell[a], Terje Gobakken[c], Erik Næsset[c], Ross F. Nelson[d], Ronald E. McRoberts[e], Timothy G. Gregoire[f], Göran Ståhl[a]

[a]Department of Forest Resource Management, Swedish University of Agricultural Sciences, SLU Skogsmarksgränd, Umeå SE-90183, Sweden
[b]USDA Forest Service, Pacific Northwest Research Station, USA
[c]Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O.Box 5003, Å NO-1432s, Norway
[d]NASA/Goddard Space Flight Center, Greenbelt, MD20771, USA
[e]USDA Forest Service, Northern Research Station, USA
[f]School of Forestry and Environmental Studies, Yale University, New Haven, CT, USA

ARTICLE INFO

Article history:
Received 27 April 2016
Received in revised form 26 January 2017
Accepted 1 February 2017
Available online xxxx

Keywords:
Airborne laser scanning
Design-based inference
Monte Carlo simulation
Variance estimation

ABSTRACT

Forest resource assessments utilizing remotely sensed auxiliary data are becoming increasingly important due to their ability to provide precise estimates of forest parameters at low cost. In presenting results from such surveys, it is important to provide not only estimates of the target parameters, but also their confidence intervals, which provide the range of values wherein the true value is located with a certain level of confidence. If such an interval is narrow the point estimates from the survey can be considered very reliable. In estimating the confidence interval the variance of an estimator must first be estimated. Unbiasedness, i.e. that an estimator on average coincides with the true value, is an important property also for variance estimators. Another important property is that the variance estimator itself has low variance, not least in cases when the variance estimates obtained with the estimator may not be strictly positive. One such important case is when two-stage designs are used to first allocate sample clusters in the form of strips from which auxiliary data, such as metrics derived from airborne laser scanning, are obtained; field data are then derived from sample plots beneath each sample strip in a second stage. In this article we compare two variance estimators for such surveys. The first estimator is a standard estimator suggested in reference textbooks on model-assisted sampling. The second estimator is proposed by the authors, and utilizes the auxiliary data to a larger extent. Through Monte Carlo simulation we show that both variance estimators are approximately unbiased, but that the new estimator is more stable (i.e., has lower empirical variance) and provides empirical confidence interval coverage rates that coincide more closely with the nominal coverage rates.

1. Introduction

The demands for forest resource assessments are increasing (e.g., Tomppo et al., 2010). An important driver of this development is the increased focus on renewable resources, including the substitution of fossil fuels by biofuels, and the importance of forests for preserving biodiversity. Formal commitments are expressed in agreements such as the United Nations Framework Convention on Climate Change, and its related protocols and mechanisms (e.g., IPCC, 2007).

For a long time, surveys of forest resources in the northern hemisphere have mainly been conducted through field inventories (Tomppo et al., 2010). However, an important current trend is that data acquired through remote sensing are utilized as auxiliary information (e.g.,Tomppo et al., 2008; Gregoire et al., 2011). Lately, airborne laser scanning (ALS) of strips combined with field sampling beneath the strips has been applied in several studies (e.g.,Andersen et al., 2011; Gobakken et al., 2012; Næsset et al., 2013a; Saarela et al., 2015; 2016), since this type of design combined with model-assisted estimation (Särndal et al., 1992) has been shown to be efficient compared to standard field surveys (e.g., Ene et al., 2012; 2013). Several of the surveys have focused on biomass assessment as a basis for estimating greenhouse gas emissions.

In reporting the results from the above type of surveys, as well as from other surveys, it is important to provide not only point

* Corresponding author.
E-mail address: svetlana.saarela@slu.se (S. Saarela).

http://dx.doi.org/10.1016/j.rse.2017.02.001
0034-4257/

estimates of the target parameters but also the corresponding variance estimates, since these provide a means to construct confidence intervals for the target parameters. With narrow confidence intervals, the reported point estimates can be considered trustworthy in the subsequent reporting and decision making while the opposite holds for wide intervals. For example, uncertainty assessments are mandatory for greenhouse gas emissions reporting according to the good practice guidance developed by the Intergovernmental Panel on Climate Change (IPCC) (IPCC, 2007). However, the variance estimator is a random variable and can produce small values (or even negative values). Thus, we can have narrow intervals that mislead us to think that the estimate is good while in fact the variance estimator has a large variance. This is why it is important, not only that the variance estimators are unbiased, but also that they are precise, so that they provide trustworthy estimates of the actual uncertainty linked to an emissions estimate.

In this study we address variance estimation in connection with two-stage sampling designs combining, e.g., ALS and field data. We developed a new, prediction-based, model-assisted variance estimator and compared it with the Horvitz-Thompson (HT) model-assisted variance estimator traditionally used in connection with two-stage sampling designs (e.g., Särndal et al., 1992; Gregoire et al., 2011). Specifically, we assessed the bias and variance of the traditional and the new variance estimator, as well as the empirical coverage of confidence intervals (CIs) provided by the two variance estimators. The studies were conducted through repeated sampling of a simulated population. A precondition for using our variance estimator is that auxiliary data are available for the first stage sampling units (e.g., ALS strips), which is typically the case in the type of surveys addressed.

We denote the proposed variance estimator "prediction-based, model-assisted" since it utilizes the available auxiliary information for improving the precision of the estimator, in contrast to the traditional HT model-assisted variance estimator (e.g., Särndal et al., 1992) used for this type of design.

2. Material and methods

2.1. Simulated study population resembling an area in Northern Sweden

Two-stage sampling can be considered a special case of two-phase sampling, where the subsampling within primary sampling units (PSUs; clusters in the form of strips) is made independently among clusters and invariantly of any observations made in the clusters (Särndal et al., 1992). In our study the PSUs are strips, tessellated into secondary sampling units (SSUs), grid-cells of size 17.73 m × 17.73 m, which correspond in size to a circular field plot of 10 m radius; along the PSU strips ALS data are collected so that ALS metrics are obtained from each SSU grid-cell. Growing stock volume (GSV) values are obtained from random samples of SSU grid-cells beneath each PSU strip.

Specifically for this study we created a 314 km^2 large simulated population, resembling the forest conditions in the county of Västerbotten in Northern Sweden. A multivariate copula distribution (Nelsen, 2006) was employed. As reference data we used Swedish national forest inventory (NFI) data collected from 77 circular plots of 10 m radius during summer 2009 from the Västerbotten county. GSV values per hectare were calculated for each field plot based on the field measurements (Fridman et al., 2014). ALS metrics for each NFI plot were calculated using the FUSION software (McGaughey, 2012), based on data from a nationwide ALS survey performed in Sweden, mainly for producing a new digital terrain model but also for mapping forest resources. The ALS density in this survey was about 0.5 returns m^{-2}, the data were collected during leaf-off period in 2009. Minimum and maximum height ($h_{\min}$, $h_{\max}$), skewness and kurtosis (h_{skew}, h_{kurt}), and the 30th, 50th and 60th percentile of the height values distribution (h_{P30}, h_{P50}, h_{P60}) of the first returns point cloud were selected for the modelling of GSV based on ALS data, using generalized linear modelling (GLM) with a log link and the Akaike information criterion for selecting the appropriate number of predictor variables. Only plots with non-zero values of GSV were used. For purpose of illustration, the parameter estimates of the model built based on the data from the 77 NFI plots, their standard errors, corresponding *p*-values and the pseudo-coefficient of determination (R^2_{pseudo}) (McRoberts et al., 2016) are presented in Table 1. However, note that during the Monte Carlo simulations the model parameters were estimated each time new sample was selected.

For simplicity in this methodological study, and to avoid any potential difficulties in interpreting the results due to differences in forest area among the strips, non-forest areas (mires, rock outcrop, small lakes, and agricultural lands; all occurring sparsely in the study area) were treated as forest and assigned values from the most similar forest plot based on Euclidean distances derived from the Landsat spectral values.

For each NFI plot we also extracted Landsat Enhanced Thematic Mapper Plus (ETM+) spectral values of blue, green, red, near infrared (IR), and two shortwave IR bands from a scene acquired on March 28, 2009 (WRS path and row 193/015), using the bilinear interpolation technique. Before the values were extracted, the Landsat scene was geo-rectified to the SWEREF99–TM geographical coordinate system. To assign Landsat ETM+ values to each SSU grid-cell, the original Landsat 30 m × 30 m pixel data were resampled using the "cubic convolution" method.

Based on the reference dataset a population of 5 million hypothetical grid-cells was generated employing the canonical vine (C-vine) copulas modelled with the packages "VineCopula" (Schepsmeier et al., 2015) and "CDVine" (Brechmann and Schepsmeier, 2013) for the statistical software R (R Core Team, 2015). Q-Q graphs comparing quantiles of (marginal) reference data distributions with quantiles of (marginal) simulated distributions for GSV and ALS variables are presented in Fig. 1. Put simply, we used the copula technique to simulate a large dataset from our smaller empirical reference dataset, so that the multivariate distribution of the simulated data corresponds to the multivariate distribution of the empirical reference data (Nelsen, 2006). Each actual grid-cell in the study area was populated with GSV, ALS metrics, and Landsat data from the 5 million observations using nearest neighbour imputation, with Euclidian distances based on the Landsat spectral values as a link. We divided the study area into 80 strip (PSUs), each containing 12 500 grid-cells (SSUs). Thus, the study region was rectangular with a length of 2000 grid-cells and a width of 500 grid-cells (Fig. 2), i.e. an area of 314 km^2.

In addition to the simulated forest population we performed a simple but more generic study where one million population units were simulated and divided into 80 clusters (PSUs). In this

Table 1
Estimated model parameters, their standard errors (SE) and *p*-values, and the R^2_{pseudo}. The dependent variable is the natural logarithm of GSV. The model was built using data from 77 NFI plots.

Variable	Estimated model parameter	SE	*p*-value
Intercept	3.49	0.23	0.00
$h_{\min}$	0.29	0.06	6.11×10^{-06}
$h_{\max}$	0.18	0.04	1.31×10^{-04}
h_{skew}	−1.01	0.29	7.72×10^{-4}
h_{kurt}	−0.1	0.04	3.02×10^{-2}
h_{P30}	−0.15	0.05	9.08×10^{-3}
h_{P50}	0.49	0.16	3.14×10^{-3}
h_{P60}	−0.52	0.17	3.10×10^{-3}
			$R^2_{pseudo} = 0.75$