Remote Sensing of Environment xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume

Johannes Breidenbach [a],*, Ronald E. McRoberts [b], Rasmus Astrup [a]

[a] *Norwegian Institute of Bioeconomy Research (NIBIO), Postboks 115, 1431 Ås, Norway*
[b] *USDA Forest Service, Forest Inventory and Analysis, St. Paul, MN, United States*

ARTICLE INFO

Article history:
Received 22 February 2015
Received in revised form 7 July 2015
Accepted 18 July 2015
Available online xxxx

Keywords:
Forest inventory
Model-based inference
Synthetic estimator
Variance estimation
Image matching

ABSTRACT

Due to the availability of good and reasonably priced auxiliary data, the use of model-based regression-synthetic estimators for small area estimation is popular in operational settings. Examples are forest management inventories, where a linking model is used in combination with airborne laser scanning data to estimate stand-level forest parameters where no or too few observations are collected within the stand. This paper focuses on different approaches to estimating the variances of those estimates. We compared a variance estimator which is based on the estimation of superpopulation parameters with variance estimators which are based on predictions of finite population values. One of the latter variance estimators considered the spatial autocorrelation of the residuals whereas the other one did not. The estimators were applied using timber volume on stand level as the variable of interest and photogrammetric image matching data as auxiliary information. Norwegian National Forest Inventory (NFI) data were used for model calibration and independent data clustered within stands were used for validation. The empirical coverage proportion (ECP) of confidence intervals (CIs) of the variance estimators which are based on predictions of finite population values was considerably higher than the ECP of the CI of the variance estimator which is based on the estimation of superpopulation parameters. The ECP further increased when considering the spatial autocorrelation of the residuals. The study also explores the link between confidence intervals that are based on variance estimates as well as the well-known confidence and prediction intervals of regression models.

1. Introduction

The use of airborne laser scanning (ALS) data in operational forest management inventories (FMI) has a long tradition in the Nordic countries (Maltamo & Packalen, 2014; Næsset, 2014). Usually, the area-based approach (ABA) is adopted, where the study area is gridded into small cells for which height and density metrics are calculated from ALS data (Næsset, 1997, 2014). A model linking the variable of interest, such as timber volume, to the ALS metrics is estimated using field sample plots where the variable of interest and the ALS data are both available. The linking model is then applied to the grid cells to map the timber volume. A main product of an FMI is a map of mean stand-level timber volume where the mapped value for each stand is calculated as the mean of timber volume predictions for grid cells whose centers are in the stand.

Although the ABA was developed using ALS data, it is well suited for the use with other remote sensing data providing high-resolution height information. For example, photogrammetric image matching data are increasingly popular to estimate forest parameters using the ABA due to the increasing availability of high-quality digital terrain models as well as improved hard- and software (e.g., Bohlin, Wallerman, & Fransson, 2012; Breidenbach & Astrup, 2012; Vastaranta et al., 2013).

In the terms of survey sampling, the ABA is one form of small area estimation (SAE), since the stands are so small or so remote that few if any sample plots are located within them. From the perspective of SAE, ALS metrics are auxiliary data and aggregating the predictions for grid-cells at stand-level is a synthetic estimate for a small area (Rao, 2003, p. 46). This estimate is termed synthetic because only model predictions are used, with no correction for model prediction errors. While design-based estimators are generally preferred in forest inventories if enough field observations are available because they are asymptotically unbiased, synthetic estimators are generally model-based (Chambers & Clark, 2012, p. 169). The basic difference between model-based and design-based inference is the source of randomness (Kangas, 2006). Whereas randomness is introduced by sample selection in design-based inference and observations are assumed to be fixed values, the observations are assumed to be a random realization of a joint distribution known as the superpopulation in model-based inference. One consequence of the differences in these underlying assumptions is that probability samples are not necessary for model-based estimators.

* Corresponding author.
E-mail address: Johannes.Breidenbach@nibio.no (J. Breidenbach).

http://dx.doi.org/10.1016/j.rse.2015.07.026
0034-4257/

For an introduction to model-based inference and a comparison to design-based inference see Gregoire (1998).

Although studies aiming at small area estimation using remotely sensed data are plentiful, the uncertainty of estimates is often ignored for management applications. While the number of SAE studies in forestry including inference is increasing (e.g., Breidenbach & Astrup, 2012; Goerndt, Monleon, & Temesgen, 2013; Lappi, 2001; Magnussen, Mandallaz, Breidenbach, Lanz, & Ginzler, 2014; Steinmann, Mandallaz, Ginzler, & Lanz, 2013), the number of studies that provide methods for stand-level inference using synthetic estimators is small (e.g., Kangas, 1996; Mandallaz, 2013; McRoberts, 2006).

In this study, we focus on synthetic estimation, which is relevant for small areas that frequently contain no or too few observations to apply other estimators. The context of the study is model-based inference which assumes that an entire distribution of observations is possible for each population unit. In this context, prediction of an individual observation (a finite population value) is distinguished from estimation of the expected value of the distribution of observations (a superpopulation parameter) (Kangas, 2006, p. 40). Although the prediction of an observation and the estimate of its expected value are the same for models relevant in our context, the variance estimates may be quite different. This paper focuses on different approaches to estimating the variances.

A variance estimator based on the estimation of superpopulation parameters just considers the variance resulting from the estimation of the model parameters and is therefore independent of stand size (e.g., Kangas, 1996; Mandallaz, 1991; McRoberts, Andersen, & Næsset, 2014).

Kangas (2006) and McRoberts (2006) described a variance estimator which is based on predictions of finite population values rather than estimates of superpopulation parameters. The estimated variance is therefore dependent on stand size. Kangas (2006) described the basic form of the variance estimator in a general setting, not specifically for SAE. McRoberts (2006) extended the variance estimator for spatial autocorrelation and applied it to the binary variable forest/non-forest. We modify the variance estimator described by McRoberts (2006) for application to a continuous response variable for which we accommodate heteroskedasticity and spatial autocorrelation.

The aim of this study is to compare a variance estimator based on the prediction of superpopulation parameters with variance estimators based on predictions of finite population values in the context of synthetic estimation. Furthermore, we link the variance estimators to the concepts of prediction intervals and confidence intervals well-known from regression analysis.

In a case study, we use Norwegian National Forest Inventory (NFI) data to estimate stand-level mean timber volume. Photogrammetric image matching data processed using the ABA serve as auxiliary information. To compare estimators, the empirical coverage proportions (ECP) of the confidence intervals based on the different variance estimators are obtained using independent validation data.

2. Methods

2.1. Estimators

A linking model describes the statistical relation between the response (variable of interest) denoted y and the auxiliary variables x which, in this case, are obtained from remotely sensed data

$$y_i = f(\boldsymbol{X}_i, \boldsymbol{\beta}) + \varepsilon_i, \quad i = \{1, \ldots, n\}, \quad \boldsymbol{\varepsilon} \sim N\left(0, \sigma_\varepsilon^2 \boldsymbol{W}\right) \tag{1}$$

where i indexes observations, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \cdots, \boldsymbol{X}_n^T)^T = (1\ \boldsymbol{x}_1 \cdots \boldsymbol{x}_p)$ is a $n \times (p + 1)$ design matrix, p is the number of auxiliary (explanatory) variables, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ is a vector of model parameters to be estimated, and ε_i is a residual.

The residual variance is expressed as the product of σ_ε^2 and a $n \times n$ matrix $\boldsymbol{W}$ where σ_ε^2 is the mean square residual. In the case of homogeneous variances, $\boldsymbol{W}$ is an identity matrix ($w_{ii} = 1$). In the case of heteroskedsticity, the diagonal elements contain appropriate weights w_{ii} that result from a variance model. In the case of autocorrelation, also the off-diagonal elements contain appropriate weights w_{ij} that result from a model describing the correlation pattern among the residuals.

To simplify the following estimators, we assume a linear model f. While we assume that the auxiliary variable is available wall-to-wall in the areas of interest, the response is only observed at a sample of the population. In forest inventories, the response is typically observed at n sample plots systematically distributed over the landscape with distances between plots in the range 100–1000 m.

In general, synthetic estimators describe a group of estimators for small areas that are based on a population level model, assuming that the characteristics of the large area hold for the small areas (Gonzalez, 1973; NCHS, 1968). This means that differences in estimates for different areas are explained by differences in the auxiliary variables rather than differences in relationships between the response and auxiliary variables (Särndal, Swensson, & Wretman, 1992, p. 411). Synthetic estimators are potentially biased, but the bias can be small if the linking model holds in the small area.

Suppose, enough observations were available within a small area to support fitting a local linking model just for the small area. If the estimated model parameters of the local linking model are very similar to those for the linking model fitted to the large area, the bias of the synthetic estimator would be small. However, usually large numbers of observations within stands are not available in operational forest inventories. The bias of the synthetic estimator will therefore usually remain unknown.

The regression-synthetic estimator (Rao, 2003, p. 46), as one specific synthetic estimator, is the mean of predictions of a linking model for units within a small area. If the linking model is a linear regression model as assumed in this study, the mean of the model predictions equals the product of the means of the auxiliary variables ($\overline{\boldsymbol{X}}$) and the regression coefficient estimates

$$\widehat{\overline{Y}}_m = \overline{\boldsymbol{X}}_m^T \hat{\boldsymbol{\beta}} = \frac{1}{N_m} \sum_{i=1}^{N_m} f\left(\boldsymbol{X}_i, \hat{\boldsymbol{\beta}}\right) = \frac{1}{N_m} \sum_{i=1}^{N_m} \hat{\mu}_i \tag{2}$$

where $i = \{1, \ldots, N_m\}$, N_m is the number of population elements within small area m, and $m = \{1, \ldots, M\}$ where M is the total number of small areas. The upper case letter $\widehat{\overline{Y}}_m$ is used for the small area-level estimate which is the estimated mean of predictions for population elements $\hat{\mu}_i$. Very small areas can also consist of only one population element in which case $\widehat{\overline{Y}}_m = \hat{\mu}_i, i = 1$. The notation $\hat{\mu}$ indicates that the model prediction is an estimate of the superpopulation parameter (the expected value of the linking model given the explanatory variables), not a prediction of the observation, $\hat{y}$. The first representation of estimator (2) is applicable for linear models (Kangas, 1996; Mandallaz, 1991), the second and third representation is more generally valid (e.g., also for nonlinear models) (McRoberts, 2006; McRoberts, Næsset, & Gobakken, 2013).

In the ABA, the population elements are often designated grid-cells and the small areas are typically stands. Typically, some of the grid cells will overlap with the sample plots used to fit the linking model (1). While Næsset (1997) was among the first to apply the regression-synthetic estimator in the ABA, Kangas (1996), McRoberts (2006), Mandallaz (2013), and McRoberts et al. (2013) have, among others, described the variance of the estimator in a forest inventory setting. Because estimator (2) is the mean of the predictions, the variance is the two-dimensional mean of the covariances of the predictions

$$\widehat{Var}_p\left(\widehat{\overline{Y}}_m\right) = \overline{\boldsymbol{X}}_m^T \hat{\Sigma} \overline{\boldsymbol{X}}_m = \frac{1}{N_m^2} \sum_{i=1}^{N_m} \sum_{j=1}^{N_m} \widehat{Cov}\left(\hat{\mu}_i, \hat{\mu}_j\right) \tag{3}$$

Please cite this article as: Breidenbach, J., et al., Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume, *Remote Sensing of Environment* (2015), http://dx.doi.org/10.1016/j.rse.2015.07.026