Remote Sensing of Environment 114 (2010) 911–924

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data

Johannes Breidenbach [a,*], Erik Næsset [a], Vegard Lien [a], Terje Gobakken [a], Svein Solberg [b]

[a] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway
[b] Skog og Landskap, Postboks 115, 1431 Ås, Norway

ARTICLE INFO

Article history:
Received 31 August 2009
Received in revised form 28 November 2009
Accepted 5 December 2009

Keywords:
Forest inventory
ITC
Lidar
Digital aerial photographs
Data fusion
*k*NN
MSN
Random forest
Nonparametric methods

ABSTRACT

While forest inventories based on airborne laser scanning data (ALS) using the area based approach (ABA) have reached operational status, methods using the individual tree crown approach (ITC) have basically remained a research issue. One of the main obstacles for operational applications of ITC is biased results often experienced due to segmentation errors. In this article, we propose a new method, called "semi-ITC" that overcomes the main problems related to ITC by imputing ground truth data within crown segments from the nearest neighboring segment. This may be none, one, or several trees. The distances between segments were derived based on a set of explanatory variables using two nonparametric methods, i.e., most similar neighbor inference (MSN) and random forest (RF). RF favored the imputation of common observations in the data set which resulted in significant biases. Main conclusions are therefore based on MSN. The explanatory variables were calculated by means of small footprint ALS and multispectral data. When testing with empirical data the new method compared favorably to the well-known ABA. Another advantage of the new method over the ABA is that it allowed for the modeling of rare tree species. The results of predicting timber volume with the semi-ITC method were unbiased and the root mean squared error (RMSE) on plot level was smaller than the standard deviation of the observed response variables. The relative RMSEs after cross validation using semi-ITC for total volume and volume of the individual species pine, spruce, birch, and aspen on plot level were 17, 38, 40, 101, and 222%, respectively. Due to the unbiasedness of the estimation, this study is a showcase for how to use crown segments resulting from ITC algorithms in a forest inventory context.

1. Introduction

In the recent years, airborne laser scanning (ALS) has become an operational tool for producing forest inventory data with wall-to-wall coverage over large areas by combing the ALS data with conventional measurements from field sample plots. In the Nordic countries, particularly in Norway, most new forest inventories are based on ALS data (Næsset 2004; Næsset 2009). Basically, two methods for analyzing the ALS data can be distinguished: the area based approach (ABA) (Næsset, 2002) and the individual tree crown approach (ITC) (Leckie et al., 2003b).

In the ABA, ALS raw data are aggregated on the inventory plot level by describing them according to height and canopy density metrics (Næsset, 2002). These metrics can subsequently be used as independent variables in regression models. Likewise, the response is an aggregation of single tree measurements on sample plots and may be univariate as in the case of mean tree height (Næsset, 1997a), timber volume (Næsset, 1997b; Means et al., 2000), biomass (Nelson et al., 1988; Næsset and Gobakken, 2008) or forest fuel parameters (Andersen et al., 2005). Examples of multivariate responses estimated with parametric methods include the simultaneous estimation of height, diameter, and volume (Næsset et al., 2005), diameter distributions (e.g., Gobakken and Næsset 2004; Breidenbach et al., 2008b) or bivariate height–diameter distributions (Mehtätalo et al., 2007; Breidenbach et al., 2008a).

In most of the cases however, nonparametric k nearest neighbor (kNN) methods are applied to estimate a multivariate response such as timber volume (Maltamo et al., 2006; Packalén and Maltamo, 2007) or diameter distributions by tree species (Packalén and Maltamo, 2008). In the context of kNN, it appears expedient to denote observations where response and predictor variables are known as references. Observations for which only predictor variables are available are referred to as targets (Stage and Crookston, 2007). The 'estimation' of response variables for target units is based on imputing values of the k nearest neighboring references. If k is greater than 1, the imputed values are a (weighted) average of the k nearest neighbors for each observed response variable. Several mathematical methods exist to determine the proximity between targets and

* Corresponding author. Tel.: +47 6496 5737; fax: +47 6496 5802.
E-mail address: johannes.breidenbach@umb.no (J. Breidenbach).

0034-4257/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2009.12.004

references based on the predictor variables. Among the most common with respect to ABA is the method of most similar neighbor inference (MSN) (Moeur and Stage, 1995) and random forests (RF) (Breimann, 2001). While in MSN the distance is derived using canonical correlation analysis, an ensemble of regression trees is grown in RF and the distance between neighbors depends on how often they share the same final node.

Packalén et al. (2009) (see also Packalén (2009)) used MSN for estimating tree species specific timber volume and described a method for fusing ALS and optical data that is employed also in the present study. Nothdurft et al. (2009) combined MSN with mixed model theory that improved estimates in a hierarchical data setting. They fused ALS and stand register data to derive several tree species specific forest parameters. Hudak et al. (2008b) compared different methods to determine the proximity of neighbors and found that RF outperformed MSN with respect to root mean squared error (RMSE). In another study, Hudak et al. (2008a,c) used RF to estimate total basal area and tree density in a study area dominated by coniferous trees in Idaho, USA. With their software package which also includes ABA example data, Crookston and Finley (2008) made *k*NN techniques easily available for a larger community.

In the ITC approach, either canopy height models (CHM) (e.g., Hyyppä et al., 2001; Persson et al., 2002; Leckie et al., 2003a; Solberg et al., 2006) or the ALS raw data point clouds (Reitberger et al., 2009) are segmented into (single) tree crowns. Tree properties can subsequently be estimated using the segment properties, such as segment area or ALS-derived height metrics as explanatory variables. This method is more intuitive than ABA since the response variable refers to the single tree, which is in fact the smallest unit on which forest management is carried out. In addition, ITC provides the tree coordinates which may be of use in single tree harvesting operations and in growth predictions with distance-dependent growth models.

The segmentation errors of any segmentation algorithm can be attributed to i) missing trees, ii) segmentation of objects that are not actually trees, iii) oversegmentation of tree crowns (i.e., one crown is split into several segments), and iv) clustering of several crowns in one segment. In terms of classification accuracy assessment, errors of type i) and ii) will result in errors of omission and commission, respectively (e.g., Campbell, 2008, pp401). Oversegmentation, i.e., type iii) errors, can also be seen as a special case of type ii) errors which result in one correct segmentation and one or several errors of commission. Likewise, clustering of several crowns would result in one correct classification and one or more errors of omission. The absolute and relative frequencies of errors of the different types will depend on factors such as algorithm design, parameter settings, and complexity of the forest.

Since these errors usually do not level out, the estimate of the response variable of interest is likely to be biased if aggregated to a larger geographical unit such as a forest stand (Maltamo et al., 2004a). Most studies related to ITC have focused on the tree properties as such, like for example crown delineation (e.g., Koch et al. 2006; Solberg et al., 2006), estimation of certain single tree attributes like height and diameter (e.g., Næsset and Økland, 2002; Persson et al., 2002; Popescu et al., 2003), tree species (e.g., Brandtberg et al., 2003; Holmgren and Persson, 2004; Ørka et al., 2009) or quality characteristics (Maltamo et al., 2009). However, only few studies have addressed the use of ITC for forest inventory. Among them are Peuhkurinen et al. (2007) who compared ITC with other approaches to derive pre-harvest information on stands. They used parametric methods to estimate tree properties without differentiation of tree species. They conclude that ITC could outperform ABA and sample plot inventories despite biased results (i.e., underestimation of timber volume). Lindberg et al. (2008) as well as Flewelling (2009) used statistical methods to reduce the problems of bias related to the use of ITC in forest inventory.

More specifically, Flewelling (2009) (see also Flewelling (2008)) used probability models for tree count and other parametric methods for estimation of diameter and height on a segment basis. He added trees from individual stratum observations to aggregated results to overcome the observed underestimation. Lindberg et al. (2008) (see also Lindberg et al. (in press)) applied seemingly unrelated regression (SUR) methods to derive tree characteristics on an individual segment basis. To retrieve unbiased estimates, Lindberg et al. (2008) combined the ITC results with ABA by using Euclidean distance based on explanatory variables. Missing trees were added from the nearest neighboring (i.e., given explanatory variables) plot. Mehtätalo (2006) proposed a method to overcome biases resulting from overlapping crowns. Another approach is to combine ABA and ITC by using the first for a subdominant layer and the latter for dominant trees (Maltamo et al., 2004b).

In this study, we propose a new statistical method to handle errors that might remain a problem even with the most advanced ITC algorithms (e.g., Flewelling 2008). Our approach is an extension of the methods described by Lindberg et al. (2008) and Flewelling (2009) as it allows an unbiased prediction of the response without the additional step of using data from ABA or stratum-wise observations. Since a crown segment can contain none, one, or several trees, we call this approach "semi-ITC". In addition, we compare two frequently used methods for determining neighbors based on distances in the feature space, i.e., MSN and RF. While *k*NN methods have been used extensively in studies that employ the ABA (see above), there seems to be only one study that applied MSN to estimate attributes of selected single trees (Maltamo et al., 2009). As an extension to Maltamo et al. (2009) we predict a multivariate response and integrate the statistical models in an inventory procedure. To the very best of our knowledge this is the first study that applies RF in an ITC or segment level context.

In principle, all values that are measured during field work can be predicted using *k*NN methods. In this study, we focus specifically on total timber volume and volume of the four dominant tree species (i.e., Norway spruce, Scots pine, birch, and aspen) found in the selected study area to determine and assess the goodness of the predictions. Up to the present, deciduous tree species have not been considered separately in ALS related studies in Norway (Næsset 2009).

2. Material

2.1. Study area

The study area is located in the municipality of Aurskog-Høland, in southeast Norway (59°50′N, 11°34′E, 120–390 m a.s.l.) (Fig. 1). The managed forest under consideration is dominated by Norway spruce (*Picea abies*) and Scotch pine (*Pinus sylvestris*), which together have a proportion (based on timber volume) of 91%. Among the sampled trees, the coniferous trees also tend to be larger than the deciduous trees (Table 1). The most common deciduous tree species were birch species (*Betula spp.*) and trembling aspen (*Populus tremula*) with a proportion of 6 and 2%, respectively (Table 2). The 39 other trees sampled (see below) comprise Rowan (*Sorbus aucuparia*), different alder species (*Alnus spp.*), common juniper (*Juniperus communis*), Scotch elm (*Ulmus glabra*), and goat willow (*Salix caprea*). For convenience, the main tree species will be referred to as spruce, pine, birch, and aspen instead of using scientific names in the remainder of the text.

2.2. Field data

Field work was carried out in the dormant season of 2007–2008, i.e., 15 October 2007 to 14 April 2008. The field data consisted of measurements made on selected sample plots (Fig. 1). A purposive sampling design was chosen to get a diverse dataset. Therefore, the