Remote Sensing of Environment 117 (2012) 358–365

Contents lists available at SciVerse ScienceDirect

Remote Sensing of Environment

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

Improving k-nearest neighbor predictions in forest inventories by combining high and low density airborne laser scanning data

Johannes Breidenbach [a,b,*], Erik Næsset [b], Terje Gobakken [b]

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

ARTICLE INFO

Article history:
Received 1 April 2011
Received in revised form 7 October 2011
Accepted 8 October 2011
Available online 21 November 2011

Keywords:
Forest inventory
Area-based approach
Individual tree crown detection
Lidar
kNN
Biomass

ABSTRACT

Airborne laser scanning (ALS) acquisitions are frequently carried out in long (several kilometers) but narrow (several hundred meters) flight strips forming a block of overlapping and parallel strips. Small inaccuracies in the height measurements result from the interaction among the different parts of the ALS system and differ within each strip. Therefore, some additional strips are flown perpendicular to the main flight direction and across the first set of strips. Strip adjustment is carried out by using information from the additional flight strips. Even if low density data are acquired for the main block, due to the double coverage, the return density under the additional strips flown may be high enough for the use of individual tree crown (ITC) approaches. The study analyzed how this kind of high and low density ALS data can be combined in order to improve k-nearest neighbor (kNN) predictions. An ITC approach was followed to predict tree attributes for automatically segmented tree crowns on more than 580 artificially created sample plots under high density flight strips. No field-measured trees were available for these artificial plots. The predicted tree attributes of the artificial plots and the measured tree attributes of 154 sample plots were then aggregated to the plot level. For all sample plots, plot-level metrics were derived from low density ALS data. The metrics were used as explanatory variables in a kNN model to predict above-ground biomass. The kNN model was validated on external data.

The combined use of artificial plots and sample plots resulted in smaller relative root mean squared differences (RRMSD) (26.7%) compared to the use of the sample plots alone (28.5%). It can be concluded that the use of artificial plots improves the accuracy of kNN predictions. This is especially the case if the data range of sample plots for which predictions are sought is not adequately covered by the sample plots used to fit the kNN model. To illustrate this, approximately 30% of the sample plots with the largest and smallest biomass values were omitted. In this case, the combined use of artificial plots and sample plots resulted in a RRMSD of 29.7%, whereas the use of the sample plots alone resulted in a RRMSD of 37.4%.

1. Introduction

Due to the value of high quality information of forest resources for optimal management (Eid et al., 2004; Kangas, 2010), vegetation height data, such as those provided by airborne laser scanning (ALS), are used more and more as auxiliary variables in operational forest inventories (Hyyppä et al., 2008, 2009; Næsset, 2004). Especially in the Scandinavian countries, forest inventories based on the area-based approach (ABA) (Næsset, 2002) are carried out on a large scale (Næsset, 2009). In the ABA, ALS data are aggregated to the sample plot level by characterizing their height distribution with different metrics. These ALS metrics can subsequently be used as independent variables in regression models for predictions in areas without sample plots but with ALS coverage. Likewise, the response is an aggregation of single-tree measurements on sample plots.

The list of studies using ALS data in boreal forests is long and includes the prediction of mean heights (Næsset, 1997), stem number (Næsset et al., 2005) and biomass (Næsset & Gobakken, 2008). Many studies have also demonstrated the usefulness of ALS data in combination with the ABA outside Scandinavia (e.g., Andersen et al., 2005; Breidenbach et al., 2008; Means et al., 2000; Straub et al., 2009).

Due to the decreasing costs of high density (>1 return m^{-2}) ALS data, inventories based on individual tree crown (ITC) approaches (Hyyppä & Inkinen, 1999) have become more popular. In the ITC approach, individual tree crowns are detected in the ALS data using automated computer algorithms. Tree properties can subsequently be estimated using the detected crown properties, such as crown area or ALS-derived height. Many studies have reported high correlations between ALS measurements within tree crowns and tree-specific measurements such as volume, height and diameter (Hyyppä et al.,

* Corresponding author. Tel.: +47 64948981; fax: +47 64948001.
E-mail address: johannes.breidenbach@skogoglandskap.no (J. Breidenbach).

0034-4257/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2011.10.010

2001; Persson et al., 2002; Yu et al., 2010), and crown base height (Popescu & Zhao, 2008). The ITC approach can outperform the ABA (Peuhkurinen et al., 2007), especially if tree detection errors are correctly considered (Breidenbach et al., 2010). Approaches which accept that ALS detected crowns may be associated with either no, one, or several field measured trees can be subsumed as semi-ITC methods (Breidenbach et al., 2010).

K-nearest neighbor (kNN) approaches are popular statistical methods for predicting various attributes in forest inventories supported by remote sensing data (McRoberts, in press; McRoberts & Tomppo, 2007). Also many studies investigating the use of ALS data for forest inventories have applied kNN methods (e.g., Hudak et al., 2008; Nothdurft et al., 2009; Packalén & Maltamo, 2006). The main advantages of kNN methods are the simplicity in predicting multivariate response variables and the absence of distributional assumptions. One drawback of kNN methods is that predictions outside the range of the reference data inherently result in either underestimation or overestimation (Magnussen, Tomppo & McRoberts, 2010; McRoberts, 2009). This systematic lack of fit in the extrapolation range is a property of kNN approaches and is also known as extrapolation bias (Magnussen, Tomppo, & McRoberts, 2010). Magnussen, Tomppo and McRoberts (2010) proposed a general model-assisted method to dampen the systematic lack of fit in the extrapolation range. A further drawback of kNN methods are the gaps that occur in data if references are sparsely distributed in parts of the data range. An increase in the number of neighbors used for prediction (k) usually reduces problems associated with gaps but increases problems associated with the systematic lack of fit in the extrapolation range (McRoberts, 2009).

Today, ALS acquisitions are carried out in long (several kilometers) but narrow (several hundred meters) flight strips forming a block of overlapping and parallel strips. To adjust for small inaccuracies resulting from the interplay between the different parts of the ALS system, the flight strips are adjusted to each other by flying at least one flight strip, but frequently several flight strips, perpendicular to the others. Even if low density data are acquired for the main block, the return density under these perpendicular strips may be high enough to be used with ITC methods due to the combination of the two data sets (Hyyppä et al., 2008; Næsset et al., 2006).

The aim of the study was to develop a strategy capitalizing on the high ALS return density resulting from double flight strips to improve the accuracy of kNN predictions in areas of low density ALS data. The core of this strategy is to make predictions of high accuracy in the areas of high density ALS coverage using a semi-ITC approach (Breidenbach et al., 2010). Subsequently, the predictions are used in combination with the original sample plots for ABA predictions in areas with low density ALS data. The accuracy was tested on external data with above-ground biomass as the response variable.

2. Material

2.1. Study area and field data

The study area was located in the municipality of Aurskog-Høland, in south-east Norway (Fig. 1). The forest in the area is dominated by Scots pine (*Pinus sylvestris* L.) and Norway spruce (*Picea abies* (L.) Karst.) and covers a smooth topography with heights above sea level ranging between 120 m and 390 m. Less than 2% of the stands were dominated by birch (*Betula* spp.). Some other deciduous tree species, such as aspen (*Populus tremula*) and elm (*Ulmus* spp.), were considerably less abundant than birch.

Two field inventories were carried out in the study area. Following a purposive sampling design, 40 large sample plots were located under strips of high density ALS data (see next section) for research purposes (Fig. 1, lower right-hand side). In an operational forest inventory, 201 small sample plots were distributed according to a stratified systematic design (Fig. 1, upper right-hand side).

The operational forest inventory comprised a total of 201 circular plots, each of which was 200 m^2 in size. The sample plots were measured

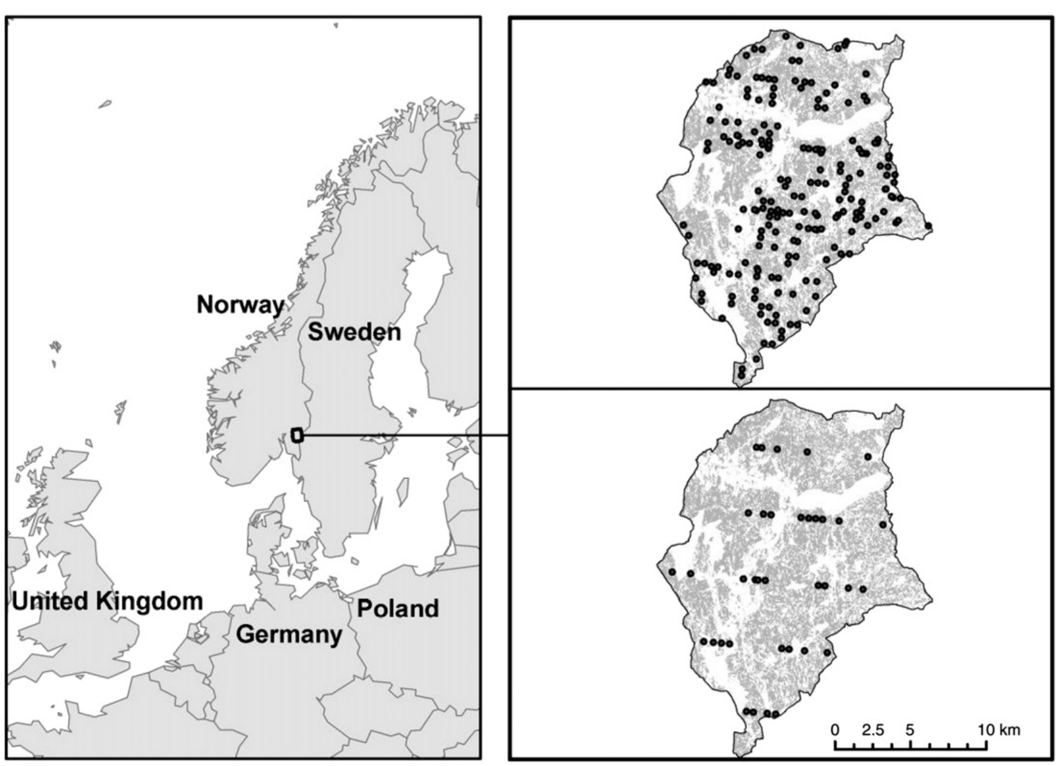

Fig. 1. Location of the study area and positions of the 201 small sample plots of the operational forest inventory (upper right-hand side) as well as 40 large sample plots (lower right-hand side). Forest areas are shown in gray in the images on the right-hand side.