Scandinavian Journal of Forest Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/sfor20

# Operational prediction of forest attributes using standardised harvester data and airborne laser scanning data in Sweden

Jon Söderberg, Jörgen Wallerman, Anders Almäng, Johan J. Möller & Erik Willén

**To cite this article:** Jon Söderberg, Jörgen Wallerman, Anders Almäng, Johan J. Möller & Erik Willén (2021) Operational prediction of forest attributes using standardised harvester data and airborne laser scanning data in Sweden, Scandinavian Journal of Forest Research, 36:4, 306-314, DOI: 10.1080/02827581.2021.1919751

**To link to this article:** https://doi.org/10.1080/02827581.2021.1919751



Published online: 05 May 2021.

SCANDINAVIAN JOURNAL OF FOREST RESEARCH
2021, VOL. 36, NO. 4, 306–314
https://doi.org/10.1080/02827581.2021.1919751

Taylor & Francis
Taylor & Francis Group



# Operational prediction of forest attributes using standardised harvester data and airborne laser scanning data in Sweden

Jon Söderberg [a]*, Jörgen Wallerman [b], Anders Almäng[c], Johan J. Möller[a] and Erik Willén[a]

[a]Skogforsk (the Forestry Research Institute of Sweden), Uppsala, Sweden; [b]Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden; [c]Sveaskog Förvaltnings AB, Ljusdal, Sweden

**ABSTRACT**

With cut-to-length harvesters, tree stems are measured and cut into different timber assortments at the time of felling. These measurement data collected from harvested trees can be used for decision-support at different levels of the forest industry chain and also for forest planning when combined with remote sensing data. The aim of this study was to examine the operational application for predicting merchantable stem volume, basal area, basal area-weighted mean tree height, basal area-weighted mean stem diameter and diameter distribution at stand level with airborne laser scanning data and harvester data from final felling operations. The area-based approach using *k*-MSN estimation was evaluated for six different variants of spatial partitioning. The results were stand level predictions with relative root mean square errors of 11–14%, 10–15%, 3–4% and 6–7% for merchantable stem volume, basal area, basal area-weighted mean tree height and basal area-weighted mean stem diameter, respectively. Predictions of stem diameter distributions resulted in error indices of 0.13–0.14. The results demonstrate that harvester data from cut forests may serve as ground truth to airborne laser scanning data and provide accurate forest estimates at stand level. The predicted diameter distributions could be useful for improving yield estimates and bucking simulations.

**ARTICLE HISTORY**
Received 7 August 2020
Accepted 15 April 2021

**KEYWORDS**
Forestry; planning; airborne laser scanning; harvester data

## Introduction

Cut-to-length (CTL) is the dominating harvesting method in Scandinavian forestry. Using CTL harvesters, tree stems are measured and cut into separate timber assortments at the time of felling, and optimisation of bucking and the subsequent flow of timber to industries is crucial in maximising revenues. Each planned harvest currently lacks sufficient information to ensure this optimisation. Möller et al. (2015) clearly identify the need for improved yield estimates, for planning as well as communication along the value chain, in the Swedish forest industry.

This study examines a possibility to capture the necessary information by using the extensive and detailed data available from previous harvests stored in data files from harvesters. Using airborne laser scanning (ALS) data, a planned harvesting area is matched to previous harvests of similar forests, and existing harvester data are then used to create a very detailed representation of the forest planned for cutting. This enables a completely new level of harvest optimisation using simulations of bucking strategies and harvesting systems (Möller et al. 2015).

Harvesters in Sweden collect data for each processed log (Arlinger et al. 2003; Rasinmäki and Melkas 2005) in accordance with the StanForD standard (Arlinger et al. 2012). In most cases, the geographical position of the machine (measured by on-board GPS mounted on the main body) is recorded for each processed tree. Harvester data in Sweden are provided daily, or even hourly, from around 1400 harvesters, transmitted through cellular networks, and stored by the forest companies or by Biometria (www.biometria.se), a data centre supporting most Swedish forest companies. These data are currently used for reporting production and controlling the timber flow (Skogforsk 2021).

Data from harvesters provide detailed information of the output in terms of length, diameter, species and timber assortment of each produced log. This information can be used to accurately reconstruct the dimensions of each harvested tree, if complemented with taper functions for estimating the length of the cut-off treetop. From this, the yield can be optimised using bucking simulations based on various price lists and assortment combinations. Upcoming harvests can be better planned and optimised if key information about the forest is available, such as stem volume, basal area, tree height, stem diameter, stem diameter distribution, tree species, stem taper and wood quality. To date, there has been little use of these data, as the data are only available for forest stands that no longer exist (Rasinmäki and Melkas 2005). For these stands, the use of harvester data combined with remote sensing data, acquired before the harvest, has been suggested.

Forest management planning in Sweden has been revolutionised by three-dimensional (3D) data about the forest

**CONTACT** Jon Söderberg jon.soderberg@skogforsk.se
*Present address: Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden

canopy captured by ALS. Accurate, large-scale maps of forest variables, available at low cost, can now support and improve decisions about silvicultural treatments compared to the subjective and manual practices used previously. Using the area-based estimation method (e.g. Næsset 2007), ALS data can provide estimates of forest variables in Scandinavian boreal forest with accuracies in terms of root mean square error (RMSE) in the range of 11–18% (in percent of true mean) for stem volume, 3–6% for basal area-weighted mean tree height and 9–13% for basal area-weighted mean stem diameter at stand level (Næsset 2007; McRoberts et al. 2010; Gobakken et al. 2014; Kukkonen et al. 2019). The accuracy produced by ALS generally outperforms traditional sources of data for management planning, such as subjective field estimation (15–25% RMSE for stem volume, and 9% RMSE for tree height), and field estimation in combination with interpretation of aerial photos viewed in stereo (15–25% RMSE for stem volume and 10% RMSE for tree height) (see, e.g. Ståhl 1988, 1992). Estimates are commonly made using non-parametric estimation methods (e.g. the *k*-MSN algorithm), also utilising a set of field-surveyed reference data plots (Maltamo et al. 2006a; Packalen and Maltamo 2007). However, acquisition of ALS data is expensive, and this has previously limited the use in operations to large forest holdings.

ALS data are now available for all forested land in Sweden, because of an ALS campaign carried out by Lantmäteriet (the Swedish National Land Survey) between 2007 and 2016, originally intended to produce a new, accurate digital terrain model (DTM) of Sweden. Updating of this initiative began in 2018 and is currently ongoing at an approximate annual scanning rate of 1/7 of the forested area in the country. Apart from the DTM, the data have also been used to produce country-wide raster maps of estimated forest variables, published online, free-of-charge for public use, and have been highly appreciated by forest owners in Sweden (Nilsson et al. 2017). These maps are produced using the area-based estimation method utilising sample plot data collected by the Swedish National Forest Inventory (NFI) (Fridman et al. 2014) as reference data. With the emergence of the detailed DTM from Lantmäteriet, a whole new level of accuracy can also be achieved with other sources of remote sensing data, such as stereo matching of high-resolution images from satellites or aerial photography, enabling 3D modelling similar to ALS (Maltamo et al. 2006b; Bohlin et al. 2012; Vastaranta et al. 2014).

ALS data enable an assessment of similarity of forest properties, and in this context data can be used to select existing harvested forest with as similar attributes as possible to those in a planned harvest site. This is clearly the most promising remote sensing data source for this application, assuming the geographical positions of harvested trees can be determined with sufficient accuracy. Most harvesters currently in operation only record GNSS coordinates of the harvester main-body, and not the harvester head. This means that if an accurate position of each harvested tree is not available, spatial aggregation of harvester tree and ALS data is necessary to form the spatial plots used as reference data.

Rasinmäki and Melkas (2005) addressed an operational application based on existing, but very limited, data (two forest stands) using a simulation approach. Various spatial divisions (in blocks, segments, raster and Thiessen polygons) of the harvested forests were used as reference plots in area-based estimation using ALS data. The method for partitioning a stand when using harvester data did not affect the accuracies of the predictions; it was the size of the plots that was significant. Generally, when using field plot data, better accuracy is expected in predictions based on smaller plot size, since each plot will contain less variation (Tuominen and Haapanen 2011), but Rasinmäki and Melkas (2005) found the inverse to be true for predictions on stand level harvester data. This is likley due to the poor accuracy of tree positions – the smaller the plots, the greater the probability that a tree will be erroneously assigned (Rasinmäki and Melkas 2005; Gobakken and Næsset 2009).

Using highly accurate, manually measured positions of each tree in combination with high-density ALS data, Holmgren et al. (2012) clearly showed the potential of a single-tree level prediction, although using data not operationally available. Saukkola et al. (2019) utilised ALS, aerial image and harvester data to predict forest variables using different positional accuracy of harvested trees and raster aggregations to form reference plots. The predictions resulted in accuracies of 25%, 25%, 6–8% and 10–11% RMSE (in percent of sampled means) for stem volume, basal area, basal area-weighted mean tree height and basal area-weighted mean stem diameter, respectively, at stand level. However, bias in predictions of basal area and stem volume was approximately 15%. Harvester head positioning produced best results using small plot sizes (254 $m^2$), while larger plot sizes were advisable if only harvester main-body positions were recorded.

Using accurate tree positioning and ALS data, Hauglin et al. (2018) produced accurate predictions of stem volume at plot level (400 $m^2$), 19–22% RMSE in high-production forest, and 32–60% RMSE in medium-production forest. Finally, highly accurate stand level predictions of stem volume and stem diameter distributions were reported by Maltamo et al. (2019), with RMSE under 9% for merchantable stem volume and error index (EI) values less than 0.2 for stem diameter distributions. This was done using ALS data and plot sizes of 200, 400, 900 and 1600 $m^2$ aggregated as raster. Performance was better for the smaller sizes, but the differences were small.

Prediction accuracy is obviously improved by high-precision positional data of each harvested tree. Boom positioning is not available from operations today, but boom angle and boom extension are already implemented in the current standard, StanForD 2010. Hence, CTL-harvester manufacturers can now relay this information from the harvester control systems. The final piece needed is machine heading, to accurately utilise the boom angle, and this could be addressed by adding a second GNSS receiver to triangulate the heading of the harvester, as described by Hauglin et al. (2017).

This study examines the operational application of estimating forest attributes and stem diameter distribution at stand level in Sweden, using tree data measured and collected from harvesters and the national ALS data, based on an area-based estimation method. Prediction accuracies of merchantable stem volume, basal area, basal area-weighted mean tree height, basal area-weighted mean stem diameter

and stem diameter distribution for final felling operations is evaluated. The influence of size and spatial form of reference plots are also considered.

## Materials and methods

The study was based on harvesting data from the forest company Sveaskog in central Sweden, and corresponding ALS data from the national laser scanning performed by Lantmäteriet. The studied area ranges from Lake Siljan (60°50′ N, 14°55′ E) in the northwest, to Lake Mälaren (59°30′ N, 16°40′ E) in the south-east (Figure 1). In this region the most common tree species are Norway spruce (*Picea abies* (L.) Karst.), Scots pine (*Pinus sylvestris* L.) and birch (*Betula* spp.).

Harvester data were provided in the StanForD XML format (Arlinger et al. 2003; Arlinger et al. 2012) and contained data from clear-felling sites only. The sites were dominated by pine or spruce and laser scanned prior to harvest. Data on 510,001 harvested trees from 168 stands with a total area of approximately 1160 ha were collected, excluding data from logging outside stands (i.e. roads and landings). From these data merchantable stem volume (solid under bark) (*V*), basal area (*B*), basal area-weighted mean tree height (*H*) and basal area-weighted mean stem diameter at breast-height (*D*) were estimated at stand level (Table 1). For *V*, measured diameter at 10 cm intervals along the stem were used and for *H*, measured log lengths were used together with estimates of tree top lengths based on stem diameter measurements (Kiljunen 2002). Each tree was assigned a geographical position, using GPS data of the harvester location at the time of felling. These operations were performed using the software hprCM developed by the Forestry Research Institute of Sweden (Skogforsk) (Siljebo et al. 2017). In a few cases, GPS data were lacking and positions for the corresponding trees were spatially interpolated from position data of trees harvested earlier and later in time, using the *pchip* algorithm from the Signal Library (R Core Team 2018). This resulted in complete tree lists, including position and information about each log. Spatial boundaries of the harvested areas were then assessed by forming a convex hull of tree positions within 25 m of each other, with a 10 m buffer added to the stand polygon border to account for positional errors and reach of the harvester boom.

The ALS were performed from altitudes between 1700 and 2300 m, with ±20° scanning angles and 20% flight path overlap, producing a density of 0.5–1 returns/m$^2$ and 0.5–0.7 m footprint (Lantmäteriet 2020). The point-cloud data were normalised to height above ground using the DTM produced by Lantmäteriet from the same point-cloud. Only points with heights above ground lower than 50 m or less than 2 m below ground were retained, and in areas of overlapping flight paths, only points from the path of steepest scan angles were retained.

The analyses were based on the area-based estimation method (Næsset 2007; Mcroberts et al. 2010), evaluating two methods of spatial aggregation of tree lists to form

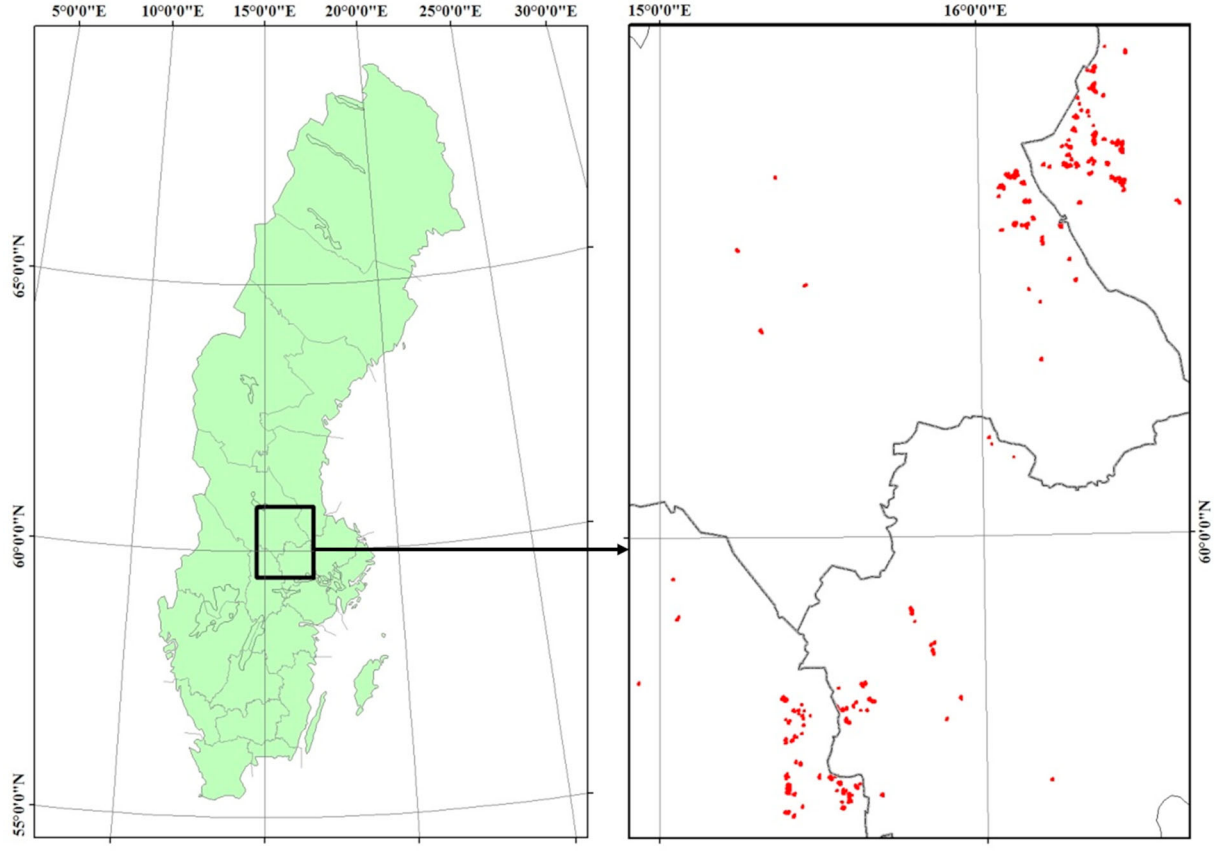


**Figure 1.** The study area in central Sweden (59°30′ N to 60°50′ N, 14°55′ E to 16°40′ E) with the harvested stands (red polygons) distributed in the four counties of Dalarna, Örebro, Västmanland and Gävleborg.

**Table 1.** Stand level forest attributes, merchantable stem volume ($V$), basal area ($B$), basal area-weighted mean tree height ($H$), basal area-weighted mean stem diameter ($D$) and number of stems ($S$), as reconstructed from harvester data and used as training and validation datasets.

| | Training Stands ($n = 88$) | | Validation Stands ($n = 80$) | |
|---|---|---|---|---|
| | min-max | mean (SD) | min-max | mean (SD) |
| $V$ ($m^3$ $ha^{-1}$) | 88–354 | 200 (47) | 118–311 | 201 (47) |
| $B$ ($m^2$ $ha^{-1}$) | 10.8–31.3 | 21.6 (3.7) | 15.3–27.9 | 21.5 (3.5) |
| $H$ (m) | 18.0–29.2 | 23.0 (2.4) | 16.6–28.9 | 23.2 (2.3) |
| $D$ (cm) | 22.5–36.0 | 28.2 (2.9) | 18.4–36.8 | 28.1 (3.1) |
| $S$ ($ha^{-1}$) | 127–800 | 462 (115) | 265–1177 | 457 (123) |

reference data plots – rasterization and segmentation. Furthermore, each method was also applied and evaluated using three different spatial sizes of reference data units, resulting in six independent estimations evaluated side-by-side. Using rasterization, each stand was partitioned using rasters of three different cell sizes, to explore the effect of positional uncertainty, with cell sides of 10, 20 and 40 m (ras10, ras20 and ras40, respectively). Segmentation was made by region merging initiated from a raster (10 m × 10 m cell size) of ALS metrics, where rectangular initial start regions were used instead of irregular Voronoi polygons (Olofsson and Holmgren 2014). The metrics used were vegetation ratio (*vr* – the proportion of returns higher than 2 m above ground), average canopy height (*ach* – the average height of first returns higher than 2 m above ground), and the 95th height percentile ($h_{95}$ – the percentile of all returns higher than 2 m above ground). Three different segmentations were made, using minimum segment sizes of 100, 300 and 900 $m^2$ (seg100, seg300 and seg900, respectively). Maximum segment size was 1,000,000 $m^2$ and merging limit was 0.1 standard deviations in all segmentations. From the complete tree lists, trees were assigned to the raster cells and segments produced in the previous steps and the corresponding variables $V$, $B$, $H$ and $D$ were calculated. Finally, ALS metrics were calculated for each raster cell and segment in terms of height percentiles ($p_{05}$, $p_{10}$, $p_{20}$, … , $p_{90}$, $p_{95}$), average height (*ah*), canopy cover (*cc*), vegetation ratio (*vr*) and height count metrics ($d_0$, $d_1$, $d_2$, $d_3$, $d_4$, $d_5$ – the percentage of points in height intervals of 2–5 m, 5–10 m, 10–15 m, 15–20 m, 20–25 m and 25–30 m, respectively). Height count metrics were also transformed to capture non-linear relations, using the natural logarithm ($lnd_i$), inversion ($invd_i$), square root ($sqd_i$) and square ($x^2d_i$), for $i = 0, \dots, 5$, corresponding to the six height intervals. Two dummy variables, *spruce* and *pine*, were classified at stand level by a proportion of merchantable stem volume of at least 70 percent for spruce and pine stem volume, respectively, and all elements were assigned these variables from the parent stand.

Prediction was based on non-parametric $k$-nearest neighbour estimation ($k$-NN), using the Most Similar Neighbour (MSN) similarity measure (Moeur and Stage 1995). This method utilises a reference dataset where the dependent variables $Y$ are known, and a set of $p$ independent variables $X$ with data available for the reference dataset as well as for all prediction points. An unobserved $Y_i$ is predicted using a weighted average of the data from the $k$ reference measurements most similar to the predicted target. In MSN, similarity is measured using a linear transformation of differences $d$ in the independent variables $X$

$$d_{ij}^2 = (x_i - x_j)A(x_i - x_j)^T \quad (1)$$

for the matrix $A$ corresponding to a given transformation of the variables $X$. In the MSN method, $A$ is defined using canonical correlation analysis of $Y$ and $X$ aiming to find the linear transformation of $X$, which explains most of the target variables $Y$, i.e.

$$A = \Gamma\Lambda^2\Gamma^T \quad (2)$$

where $\Gamma$ is the $p$ ′1 vector of canonical loadings and $\Lambda$ is the $p$ ′$p$ diagonal matrix of canonical correlation coefficients.

Using MSN distances for the $k$ most similar reference measurements, predictions were based on the weighted average of those $k$ reference measurements of $Y$, using the inverse of the MSN distance as weights (Packalen and Maltamo 2007). The canonical correlation model was fitted to the harvester measurements of $V$, $B$, $H$ and $D$ as $Y$, proportions of spruce (*spruce*) and pine (*pine*) stem volumes, and a subset of the ALS metrics, as $X$. The subset was selected by first screening the set of all metrics for collinearity using the variance inflation factor (VIF), excluding metrics with a value of 10 or more from modelling (Chatterjee and Simonoff 2013), and then applying stepwise regression of $V$, $H$ and $D$ based on *spruce*, *pine*, and the remaining ALS metrics. $V$ and $B$ were highly correlated, above 0.974, therefore $B$ was excluded from this step. To reduce edge effects from positional uncertainty and the simple stand delineation method, only elements with at least 90% of their area inside a stand polygon were included for prediction and validation. Elements with missing data or ALS metrics with less than 2 m in average height were also excluded.

The performance of predictions was evaluated by dividing the dataset in two parts, one set used as validation data, and the other used as reference data to predict $V$, $B$, $H$ and $D$ for the elements in the validation stands using the $k$-MSN method. Here, 80 stands from the 168 were randomly selected to form the validation data. Accuracy was then inferred from the predictions of the evaluation dataset using root mean square error (RMSE) (Equation (3)) and bias (Equation (4)). To validate prediction of stem diameter distributions, Reynold's error index (REI) (Reynolds et al. 1988; Gobakken and Næsset 2006) was used (Equation (5)). Here, the error index (EI), as suggested by Packalén and Maltamo (2008), was also calculated for each stand (Equation (6)). By calculating relative frequencies of stems, each stand receives an equal weight in the evaluation.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (3)$$

$$Bias = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n} \quad (4)$$

$$REI = \sum_{i=1}^{k}\frac{|f_i - \hat{f}_i|}{N}100 \quad (5)$$

$$EI = \sum_{i=1}^{k}0.5\left|\frac{f_i}{N} - \frac{\widehat{f_i}}{\hat{N}}\right| \quad (6)$$

where $\hat{y}_i$ is the predicted and $y_i$ is the observed value for stand *i*, where $i = 1, \ldots, n$, and $\bar{y}$ is the mean of all observed values. For REI and EI, *k* is the number of stem diameter classes in the stand, $f_i$ is the observed and $\widehat{f_i}$ is the predicted number of stems in stem diameter class *i*, and *N* and $\hat{N}$ are the observed and predicted number of stems for all classes, respectively.

## Results

In selection of variables for the canonical correlation model in *k*-MSN, the stepwise regression model of *V, H* and *D* based on *spruce*, *pine*, $p_{95}$, *cc*, $d_{04}$, $lnd_{01}$ and $lnd_{04}$, showed to be the most significant and were selected for constructing the canonical correlation models used in *k*-MSN. Validations of raster and segment predictions showed that merchantable stem volume was predicted with a RMSE of 26–33 $m^3\ ha^{-1}$ (11–14%), with an underestimation of 4–10 $m^3\ ha^{-1}$ (2–4%) (Table 2). For basal area, the RMSE was approximately 3 $m^2\ ha^{-1}$ (10–15%), with an underestimation of 0.2–0.8 $m^2\ ha^{-1}$ (1–3%) (Table 3). Moreover, basal area-weighted mean tree height was predicted with a RMSE less than 1 m (3–4%) regardless of aggregation method and an underestimation less than 0.2 m (<1%) (Table 4). Prediction of basal area-weighted mean stem diameter resulted in a RMSE less than 2 cm (<7%) for all spatial aggregations, with a bias of less than 0.3 cm (<1%) in all instances (Table 5). Evaluating the raster predictions, the relationship between accuracy in terms of relative RMSE and element size varied between the predicted variables, where the accuracy was approximately the same for *V*, higher for the smaller sizes for *B*, and higher for the larger sizes for *H* and *D* (Tables 2–5). These relationships between accuracy and plot size were not identical for the segment predictions, where the accuracy of *V* and *B* was higher for smaller element sizes, but not affected by size for *H* and *D*. The results from one raster variant (ras20) and one segmentation variant (seg300) are displayed in Figures 2 and 3, respectively.

Prediction of stem diameter distributions using 2 cm diameter classes and area-weighted aggregations of trees in imputed elements produced a mean relative REI of 29, 28, 30, 32, 35 and 49 for the variants ras10, ras20, ras40, seg100, seg300 and seg900, respectively (Table 6). Using the error index (EI), adapted by Packalén and Maltamo (2008), all six variants produced similar results, with a mean EI of 0.13–0.14.

**Table 2.** Validation results in terms of RMSE and bias for predictions of merchantable stem volume using three different raster cell sizes (ras10, ras20 and ras40) and three segmentation sizes (seg100, seg300 and seg900) ($n = 80$).

| | RMSE | | Bias | |
|---|---|---|---|---|
| Method | ($m^3\ ha^{-1}$) | (%) | ($m^3\ ha^{-1}$) | (%) |
| ras10 | 31.28 | 11.01 | -9.37 | -3.30 |
| ras20 | 25.88 | 11.03 | −6.74 | −2.87 |
| ras40 | 26.44 | 11.53 | −4.38 | −1.91 |
| seg100 | 27.69 | 11.60 | −9.77 | −4.10 |
| seg300 | 30.11 | 12.85 | −8.24 | −3.51 |
| seg900 | 33.17 | 14.16 | −8.00 | −3.42 |

**Table 3.** Validation results in terms of RMSE and bias for predictions of basal area using three different raster cell sizes (ras10, ras20 and ras40) and three segmentation sizes (seg100, seg300 and seg900) ($n = 80$).

| | RMSE | | Bias | |
|---|---|---|---|---|
| Method | ($m^2\ ha^{-1}$) | (%) | ($m^2\ ha^{-1}$) | (%) |
| ras10 | 3.04 | 9.99 | −0.68 | −2.23 |
| ras20 | 2.73 | 10.90 | −0.46 | −1.83 |
| ras40 | 2.74 | 11.27 | −0.26 | −1.08 |
| seg100 | 2.82 | 11.02 | −0.78 | −3.06 |
| seg300 | 3.07 | 12.29 | −0.58 | −2.34 |
| seg900 | 3.62 | 14.55 | −0.64 | −2.56 |

**Table 4.** Validation results in terms of RMSE and bias for predictions of basal area-weighted mean tree height using three different raster cell sizes (ras10, ras20 and ras40) and three segmentation sizes (seg100, seg300 and seg900) ($n = 80$).

| | RMSE | | Bias | |
|---|---|---|---|---|
| Method | (m) | (%) | (m) | (%) |
| ras10 | 0.84 | 3.64 | −0.11 | −0.47 |
| ras20 | 0.63 | 2.72 | −0.12 | −0.52 |
| ras40 | 0.62 | 2.67 | −0.07 | −0.28 |
| seg100 | 0.77 | 3.30 | −0.15 | −0.66 |
| seg300 | 0.76 | 3.25 | −0.15 | −0.65 |
| seg900 | 0.82 | 3.54 | −0.10 | −0.45 |

## Discussion

In general, the prediction method proposed in this study performed very well, producing high accuracies in terms of relative RMSE for *H* and *D* (3–4% and 6–7% RMSE, respectively), but somewhat less accurate for *V* and *B* (11–14% and 10–15% RMSE), evaluated at stand level. This with negligible bias. These results are similar to those obtained using the area-based estimation method with standard surveyed sample plots as reference data, applied on similar ALS data. In particular, the results are similar, or better, compared to the ALS-based national predictions of Sweden (Nilsson et al. 2017). The latter study was made using the same ALS data as in this study, but used field inventoried sample plots (with a plot size of 314 $m^2$) from the Swedish NFI as reference data rather than harvester data. Compared to Saukkola et al. (2019) and Hauglin et al. (2018), both of which are using harvester data as reference data, the results from the current study outperformed those obtained in both studies. However, these studies are not easily compared, since they were performed in various forest types (especially with variations in spatial homogeneity), used different harvesters, used different methods to spatially assign a harvested tree

**Table 5.** Validation results in terms of RMSE and bias for predictions of basal area-weighted mean stem diameter using three different raster cell sizes (ras10, ras20 and ras40) and three segmentation sizes (seg100, seg300 and seg900) ($n = 80$).

| | RMSE | | Bias | |
|---|---|---|---|---|
| Method | (cm) | (%) | (cm) | (%) |
| ras10 | 1.87 | 6.65 | 0.13 | 0.45 |
| ras20 | 1.58 | 5.63 | 0.15 | 0.54 |
| ras40 | 1.58 | 5.62 | 0.23 | 0.81 |
| seg100 | 1.75 | 6.23 | 0.07 | 0.26 |
| seg300 | 1.83 | 6.49 | 0.10 | 0.36 |
| seg900 | 1.95 | 6.94 | 0.17 | 0.60 |

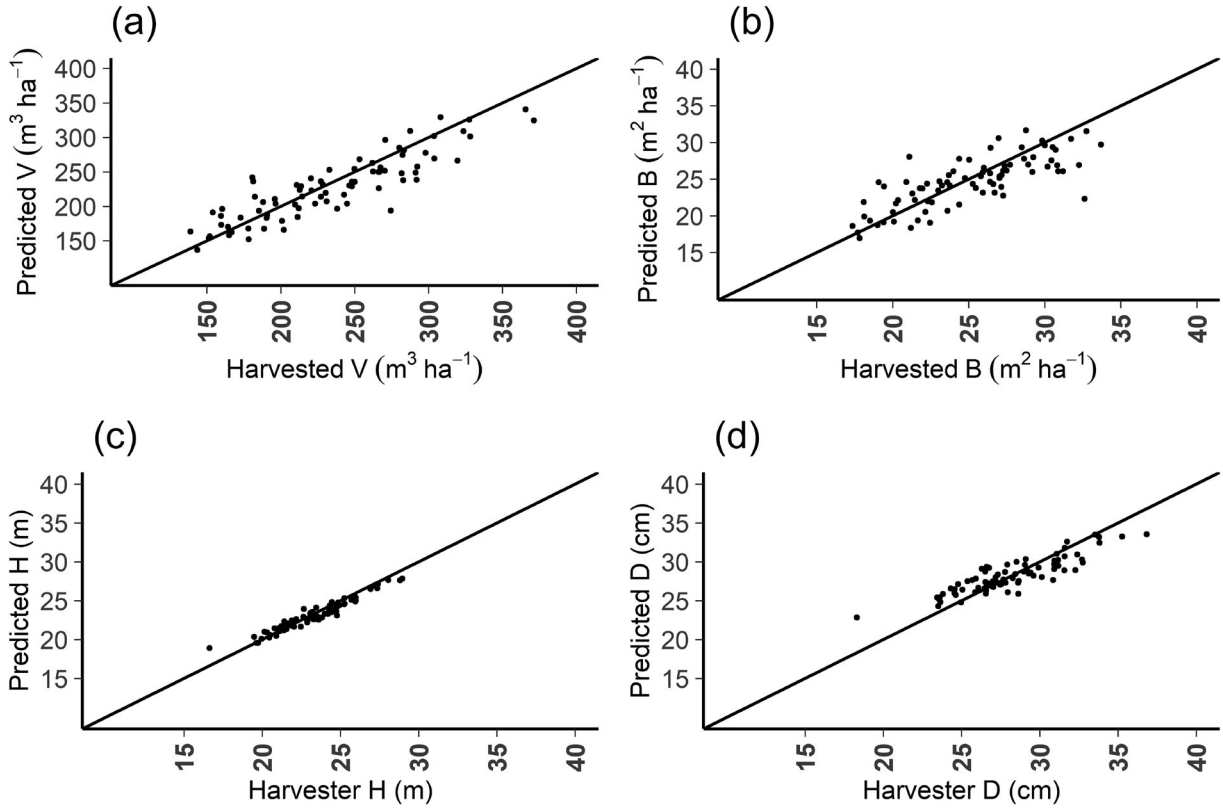


**Figure 2.** Predictions of stem volume (a), basal area (b), basal area-weighted mean tree height (c) and basal area-weighted mean stem diameter (d) compared to harvester measured dito from final felling operations, using a raster approach, with 20 m × 20 m raster cells (ras20) ($n = 80$).

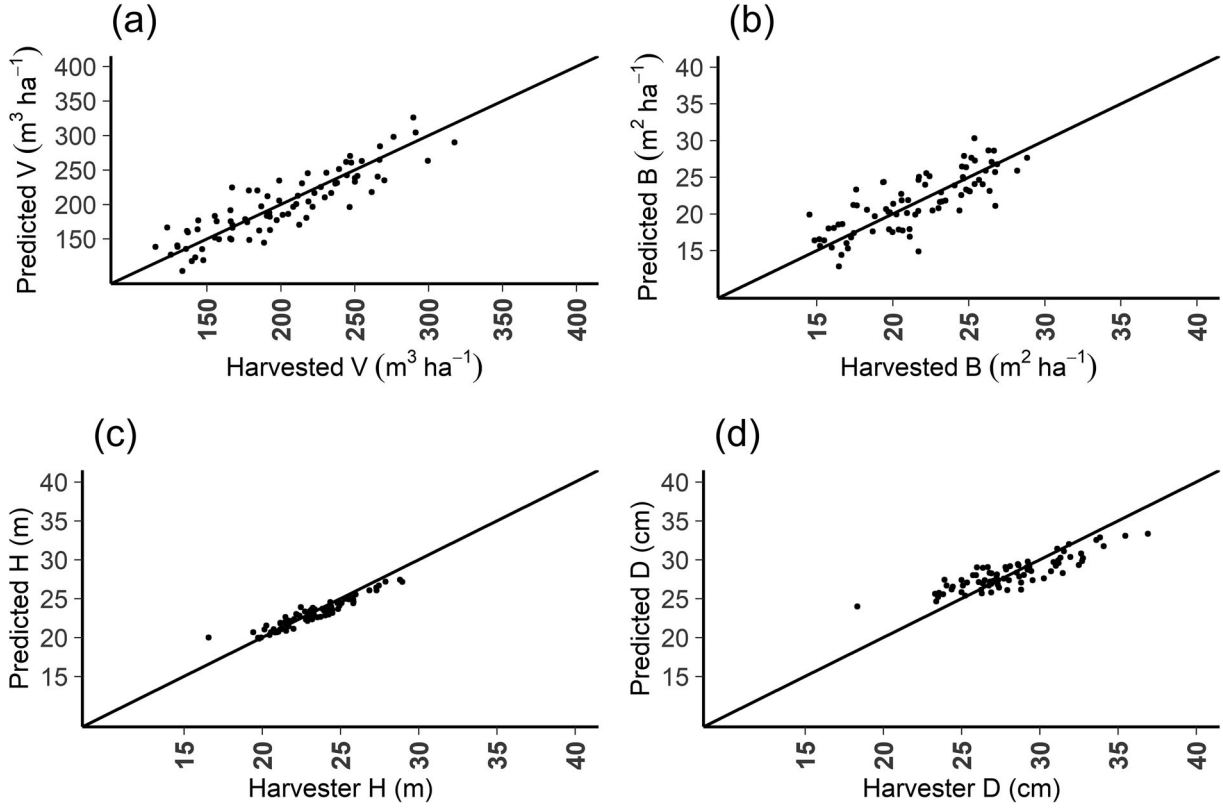


**Figure 3.** Predictions of stem volume (a), basal area (b), basal area-weighted mean tree height (c) and basal area-weighted mean stem diameter (d) compared to harvester measured dito from final felling operations, using a segmentation approach, with a minimum segment size of 300 $m^2$ (seg300).

**Table 6.** Validation results for predictions of stem diameter distributions using Reynold's error index (REI) and error index (EI) for three different raster cell sizes (ras10, ras20 and ras40) and three segmentation sizes (seg100, seg300 and seg900) ($n = 80$).

| Method | Mean REI | Min REI | Max REI | SD REI | Mean EI |
|---|---|---|---|---|---|
| ras10 | 29.37 | 11.17 | 71.90 | 11.69 | 0.14 |
| ras20 | 28.24 | 11.94 | 66.19 | 10.51 | 0.13 |
| ras40 | 29.86 | 12.59 | 62.65 | 10.24 | 0.13 |
| seg100 | 32.35 | 11.54 | 81.41 | 13.65 | 0.13 |
| seg300 | 35.39 | 12.74 | 105.13 | 18.14 | 0.13 |
| seg900 | 48.91 | 12.92 | 189.71 | 33.53 | 0.14 |

to a probable growing location, and often used very limited amounts of data.

The large dataset in this study is unique, having thousands of reference and validation elements for imputation. However, to put this study into an operational perspective, this dataset represents less than a day's worth of collected harvester data in Sweden. Data used here may be more uniform compared to other studies, with selected final felling sites dominated by coniferous trees, which may positively affect prediction accuracy. This is especially expected for *k*-NN methods; even in a limited sample of reference data observations, there are usually several similar reference data observations (close in feature space) available for each prediction target.

Evaluating predictions of stem diameter distributions, the REI for all raster-based variants were roughly equal, while for segmentation variants the fit decreased with increasing segment size. This is likely due to greatly varying size of segments within each variant, compared to equal-sized raster cells. However, when calculating the error index (EI) based on a relative distribution ignoring variation in segment size, the segmentation variants were on par with the raster variants. These results were on par or better compared to those reported by Maltamo et al. (2019), where the error indices, using 2 cm diameter classes but only evaluated for one tree species (Norway spruce), were 0.14, 0.18, 0.16 and 0.21 for 200, 400, 900 and 1600 $m^2$ cell sizes, respectively.

The different spatial aggregations of harvester data produced similar results regarding the evaluated spatial sizes, and the small differences found were not consistent between the raster and the segmentation aggregations. This is in accordance with, e.g. Saukkola et al. (2019), where raster aggregations of 761 $m^2$ cell sizes were recommended for use with harvester data where harvester head position is not available. Maltamo et al. (2019) showed a RMSE of 8–12% for predictions of *V* using raster aggregated data with 200, 400, 900 and 1600 $m^2$ cell sizes, and only noted minor improvements in accuracy for small cell sizes. In general, the raster aggregation was superior to the segmentations, although the differences were small. A segmentation approach follows the spatial variations of the forest and does produce segments of more homogeneous forests compared to raster aggregations of similar cell sizes. The lack of accurate positions of harvested trees probably negates this advantage, as harvested tree data are then often assigned to incorrect segments. On the other hand, studies using harvester head positioning to co-locate every harvested tree to ALS data more accurately do not generally show substantially higher prediction accuracy (e g., Hauglin et al. 2018; Saukkola et al. 2019).

Another interesting aspect of harvester data is the yield volumes of timber assortments, recorded for each processed log. This information could be used to predict yield volumes from harvest operations. Vähä-Konka et al. (2020) evaluated the accuracy of ALS based yield estimates from the Finnish Forest Centre by comparing to harvester measured yields and found a RMSE of 26% for total harvest removal volume, but much larger RMSEs 49–170% for individual timber assortments. Harvester data can provide large volumes of reference data for modelling yield estimates, but another vital component needed to improve yield predictions is reliable tree species composition, beyond the dominant species. This could be feasible using ALS in combination with airborne optical imaging or even a multispectral laser scanner as Kukkonen et al. (2019) has shown, or possibly even using more readily available satellite imaging data.

The methodology developed in this paper shows clear potential for large-scale applications using high-precision remote sensing data. In Sweden, the national ALS campaigns provide full cover of all forest land and can already facilitate accurate predictions of final felling outcome at any potential harvest site in Sweden. The method described in this paper can substitute yield estimates based on general functions with regional harvester data. The industry is currently adopting these new possibilities and is compiling databases of harvester data to provide reference data. Furthermore, harvester data may also complement standard plot-based field surveys as a reference data source for large-scale remote sensing-based forest mapping, although there are some issues to resolve regarding the sample representativity. In contrast to the current regime based on field plot sampling, the very large amount of harvester data available for remote sensing-based mapping tasks enables application of machine learning algorithms, such as deep neural networks (Goodfellow et al. 2016), which benefit from very large datasets. Machine learning is also expected to provide information about delicate target variables from very complex relationships in the data, direct as well as spatial, information that is not available using current methods. To some extent, it is also expected to deliver higher inference accuracy generally.

In Sweden, the centralised collection of harvester data and the national ALS campaigns provide unique possibilities to evaluate and develop the presented methodology. Large-scale mapping performance in various forest types and harvesting regimes, prediction performance of other forest variables, and integration of additional remote sensing data (e g. optical satellite image data to assess tree species) are important issues to address in future research. The ongoing improvements of positioning techniques for harvesters, especially the boom and harvester head, are expected to improve performance, as this reduces the errors introduced by spatial uncertainty. This is especially important for highly accurate single-tree level machine learning analyses using densely scanned ALS data and the extremely large quantity of data available today.

For applicability outside Sweden, national ALS data are not a requirement, the method could be applied on larger forest

holdings with a recent acquisition of ALS data. However, wide-spread use of the CTL-harvester system is a requirement, but the availability of harvester data could still face challenges of ownership rights, and concerns with privacy issues. The methodology presented here was developed for the Nordic region, and further development of algorithms to process and recreate trees from harvester measurements could be required for additional tree species. This study was limited to coniferous forest and further studies are also required to ascertain the accuracy in broad-leaved forests.

In conclusion, this study shows how harvester data from cut forests may serve as ground truth to ALS data and provide accurate forest estimates for mature, harvest ready stands. The predicted stem diameter distributions and imputed stem profiles could be useful for improving yield estimates and bucking simulations. As data from ALS are available nationwide in Sweden, it implies a possibility for operational implementation for forest companies to improve forest variable estimates, yield estimates for planning harvest operations to meet industry demand, and a feedback system for continuous improvements of data describing the forest.

## Acknowledgements

The authors would like to express gratitude to the forest company Sveaskog for providing part of the funding for this work, and to the Ljungbergs Foundation which supplied funding for the laboratory where this work has been carried out. The manuscript was compiled as part of the Mistra Digital Forest programme. The authors also express gratitude to the developers and maintenance team of the software R (R Core Team 2018), which has been extensively used for performing the analyses.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

*Jon Söderberg* http://orcid.org/0000-0001-9401-1410
*Jörgen Wallerman* http://orcid.org/0000-0002-9996-1447

## References

Arlinger J, Möller JJ, Sondell J. 2003. A description of prifiles – background, structure and examples. Arbetsrapport. Uppsala: Skogforsk.

Arlinger J, Nordström M, Möller JJ. 2012. Stanford 2010 modern kommunikation med skogsmaskiner – Stanford 2010 modern communication with forest machines. Arbetsrapport 784. Uppsala: Skogforsk.

Bohlin J, Wallerman J, Fransson JES. 2012. Forest variable estimation using photogrammetric matching of digital aerial images in combination with a high-resolution dem. Scandinavian J For Res. 27:692–699.

Chatterjee S, Simonoff JS. 2013. Handbook of regression analysis. Somerset, NJ, USA: John Wiley & Sons.

Fridman J, Holm S, Nilsson M, Nilsson P, Ringvall A, Ståhl G. 2014. Adapting national forest inventories to changing requirements – the case of the Swedish national forest inventory at the turn of the twentieth century. Silva Fenn. 48(3): article id. 1095, 1–29.

Gobakken T, Bollandsås OM, Næsset E. 2014. Comparing biophysical forest characteristics estimated from photogrammetric matching of aerial images and airborne laser scanning data. Scandinavian J For Res. 30:73–86.

Gobakken T, Næsset E. 2006. Estimation of diameter and basal area distributions in coniferous forest by means of airborne laser scanner data. Scandinavian J For Res. 19:529–542.

Gobakken T, Næsset E. 2009. Assessing effects of positioning errors and sample plot size on biophysical stand properties derived from airborne laser scanner data. Can J For Res. 39:1036–1052.

Goodfellow I, Bengio Y, Courville A. 2016. *Deep learning*. Cambridge, MA: MIT Press.

Hauglin M, Hansen EH, Næsset E, Busterud BE, Gjevestad JGO, Gobakken T. 2017. Accurate single-tree positions from a harvester: a test of two global satellite-based positioning systems. Scandinavian J For Res. 32:774–781.

Hauglin M, Hansen E, Sørngård E, Næsset E, Gobakken T. 2018. Utilizing accurately positioned harvester data: modelling forest volume with airborne laser scanning. Can J For Res. 48:913–922.

Holmgren J, Barth A, Larsson H, Olsson H. 2012. Prediction of stem attributes by combining airborne laser scanning and measurements from harvesters. Silva Fenn. 46:227–239.

Kiljunen N. 2002. Estimating dry mass of logging residues from final cuttings using a harvester data management system. Int J For Eng. 13:17–25.

Kukkonen M, Maltamo M, Korhonen L, Packalén P. 2019. Comparison of multispectral airborne laser scanning and stereo matching of aerial images as a single sensor solution to forest inventories by tree species. Remote Sens Environ. 231: article id. 111208, 1–10.

Lantmäteriet. 2020. Lantmäteriet; [cited 2020 June 23]. Available from: https://www.lantmateriet.se/sv/Kartor-och-geografisk-information/geodataprodukter/produktlista/laserdata-nh/

Maltamo M, Eerikäinen K, Packalén P, Hyyppä J. 2006a. Estimation of stem volume using laser scanning-based canopy height metrics. Forestry: An Int J For Res. 79:217–229.

Maltamo M, Hauglin M, Naesset E, Gobakken T. 2019. Estimating stand level stem diameter distribution utilizing harvester data and airborne laser scanning. Silva Fenn. 53(3): article id. 10075, 1–19 .

Maltamo M, Malinen J, Packalén P, Suvanto A, Kangas J. 2006b. Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. Can J For Res. 36:426–436.

McRoberts RE, Cohen WB, Næsset E, Stehman SV, Tomppo EO. 2010. Using remotely sensed data to construct and assess forest attribute maps and related spatial products. Scandinavian J For Res. 25:340–367.

Moeur M, Stage AR. 1995. Most similar neighbor: An improved sampling inference procedure for natural resource planning. For Sci. 41:337–359.

Möller JJ, Nordström M, Arlinger J. 2015. Förbättrade utbytesprognoser – improved yield forecasts. Arbetsrapport 880. Uppsala: Skogforsk.

Nilsson M, Nordkvist K, Jonzén J, Lindgren N, Axensten P, Wallerman J, Egberth M, Larsson S, Nilsson L, Eriksson J, Olsson H. 2017. A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the national forest inventory. Remote Sens Environ. 194:447–454.

Næsset E. 2007. Airborne laser scanning as a method in operational forest inventory: status of accuracy assessments accomplished in Scandinavia. Scandinavian J For Res. 22:433–442.

Olofsson K, Holmgren J. 2014. Forest stand delineation from lidar point-clouds using local maxima of the crown height model and region merging of the corresponding voronoi cells. Remote Sens Lett. 5:268–276.

Packalen P, Maltamo M. 2007. The k-msn method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sens Environ. 109:328–341.

Packalén P, Maltamo M. 2008. Estimation of species-specific diameter distributions using airborne laser scanning and aerial photographs. Can J For Res. 38:1750–1760.

Rasinmäki J, Melkas T. 2005. A method for estimating tree composition and volume using harvester data. Scandinavian J For Res. 20:85–95.

R Core Team. 2018. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Reynolds MR, Burk TE, Huang W-C. 1988. Goodness-of-fit tests and model selection procedures for diameter distribution models. For Sci. 34:373–399.

Saukkola A, Melkas T, Riekki K, Sirparanta S, Peuhkurinen J, Holopainen M, Hyyppä J, Vastaranta M. 2019. Predicting forest inventory attributes using airborne laser scanning, aerial imagery, and harvester data. Remote Sens (Basel). 11: article id. 797, 1–15 .

Siljebo W, Möller JJ, Hannrup B, Bhuiyan N. 2017. Hprcm – modul för beräkning av trädegenskaper och skogsbränslekvantiteter baserat på skördardata – hprcm – module for using harvester data to calculate tree properties and forest fuel quantities. Arbetsrapport 944. Uppsala: Skogforsk.

Skogforsk. 2021. StanForD [Online]. Skogforsk; [cited 2021 Feb 17]. Available from: https://www.skogforsk.se/english/projects/stanford/

Ståhl G. 1988. Noggrannheten i skogliga data insamlade med subjektiva inventeringsmetoder. Umeå, Sweden: Department of Biometry and Forest Management.

Ståhl G. 1992. A study on the quality of compartmentwise forest data acquired by subjective inventory methods [Licentiate Licentiate thesis]. Swedish University of Agricultural Sciences.

Tuominen S, Haapanen R. 2011. Comparison of grid-based and segment-based estimation of forest attributes using airborne laser scanning and digital aerial imagery. Remote Sens (Basel). 3:945–961.

Vähä-Konka V, Maltamo M, Pukkula T, Kärhä K. 2020. Evaluating the accuracy of ALS-based removal estimates against actual logging data. Ann For Sci. 77: article id. 84, 1–11.

Vastaranta M, Wulder MA, White JC, Pekkarinen A, Tuominen S, Ginzler C, Kankare V, Holopainen M, Hyyppä J, Hyyppä H. 2014. Airborne laser scanning and digital stereo imagery measures of forest structure: comparative results and implications to forest mapping and inventory update. Can J Remote Sens. 39:382–395.