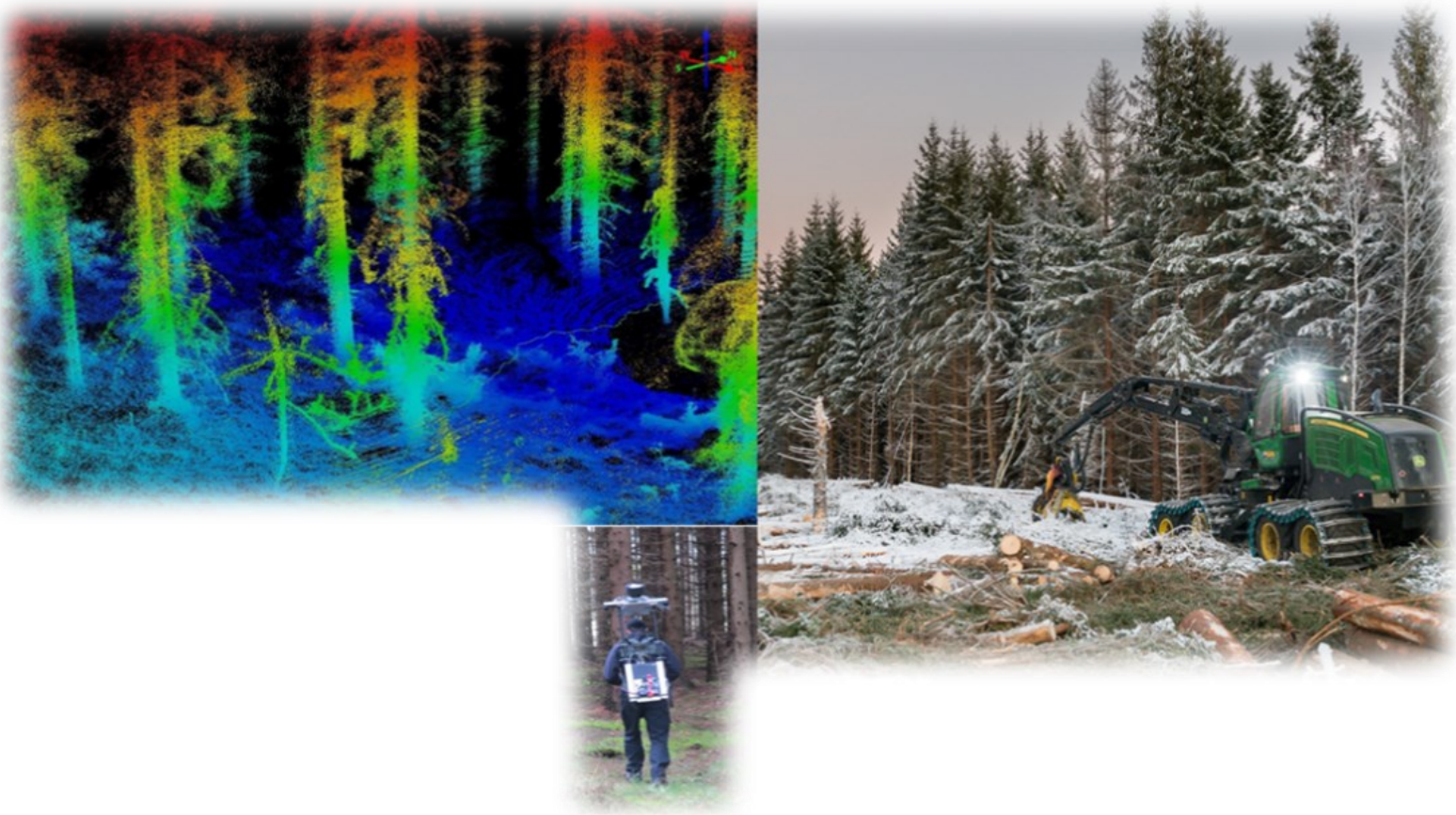
ARBETSRAPPORT 1103–2021

Volume and value recovery predictions by combining tree lists from a harvester stem database and estimated diameter distributions from a mobile laser scanner system

Liviu Theodor Ene; Jon Söderberg; Johan Möller



skogforsk

Contents

Uppsala Science Park, 751 83 Uppsala
skogforsk@skogforsk.se
skogforsk.se

Kvalitetsgranskning (Intern peer review) har genomförts i maj 2021 av
Maria Nordström, biträdande programchef. Därefter har Magnus Thor, Forskningschef,
granskat och godkänt publikationen för publicering den 30 november 2021.

Redaktör: Leslie Walke, leslie@communicaid.se
 ISSN 1404-305X

Preface

The study investigates the use of mobile laser scanning data as a sampling tool for retrieving diameter distributions to support product recovery predictions. We thank Holmen Skog and Sveaskog for providing the harvester, and Dr. Johan Holmgren (Swedish University of Agricultural Sciences, Department of Forest Resource Management) for providing the mobile laser scanning data and participating in useful discussions. The study was funded by the FORMAS project Data Assimilation (reference 2015-00063).

Uppsala, 3 March 2021

Liviu Theodor Ene, Jon Söderberg, Johan Möller & Erik Willén

Summary

Accurate product recovery predictions are necessary to enable efficient planning of harvesting operations and optimisation of wood flow to industry. Swedish forest companies and Skogforsk have therefore developed an extensive database infrastructure comprising Cut-To-Length (CTL) harvester production files. The system allows imputation of tree lists retrieved from the harvester data to stands scheduled for harvest in operational planning. Based on the imputed tree lists, the value recovery in such areas can be estimated, using either bucking simulations or, more expeditiously, price stem models.

The novelty proposed in this study was to improve the existing imputation approach, using DBH information produced by a mobile, backpack-mounted laser scanning (MLS) system in pre-harvest inventories on five experimental, mixed coniferous forest areas of 0.16 to 0.32 ha located in north-eastern Sweden. The system integrates a NovAtel SPAN-IGM-S1 Inertial Navigation System and a Velodyne VLP-16 scanner for collecting 3D point cloud data that provides information on DBH and positions for individual trees.

The working hypothesis of the study was that the information provided by the MLS system can be used to impute tree lists that are a better match to the ground-truth in terms of DBH distributions, and thereby improve the accuracy of volume predictions and value recovery. The main objective of the study was to test the working hypothesis under several scenarios, where the value recovery predictions were run using various combinations of auxiliaries.

The results indicate that:

- injecting MLS data as an auxiliary into the imputations considerably improves the match between the DBH values in the imputed and ground-truth tree lists;
- uncertainty in value recovery was reduced by half;
- using data augmentation considerably improved the results.

Conclusions:

- The limited validation dataset restricts the generalisation power of the study. However, there is evidence that DBH information retrieved from MLS measurements may have a positive effect on the imputation results, assuming that the tree species composition is known or accurately predicted.
- The augmentation method developed for this study has potential to increase the imputation accuracy, especially for value recovery, provided that detailed spatial information is available for the pre-harvest inventory data. MLS is compatible with this approach, if acquired using probability sampling schemes.
- The costs of the MLS survey were not quantified in the assessment. For operational deployment of the method, inventory costs should be considered when selecting the most cost-efficient pre-harvesting approach.
- The results opened for future research directions focusing on:
    - inclusion of external stem quality descriptors extracted from MLS data (such as straightness/sweep, branch/knot size and distribution, bark and decay) in the imputation workflow for value and product and value recovery predictions;
    - incorporation of other types of auxiliaries, such as airborne laser scanning data and/or satellite imagery. For example, textural descriptors for the forest canopy surface derived from airborne laser scanning data relate to the horizontal forest vegetation distribution, and this may replace, at least partially, the need for pre-harvest MLS inventories for DBH data collection, while satellite imagery can support the prediction of tree-species proportions required for value and product recovery calculations.

Sammanfattning

I arbetet för att stödja en effektiv planering av skördeoperationer och optimering av råvaruflödet till industrin är det nödvändigt med noggranna prognoser för produktvolymer. För att möta detta krav har Skogforsk tillsammans med svenska skogsföretag utvecklat en omfattande databasinfrastruktur som innehåller sammanställda produktionsfiler från skördare. Systemet gör det möjligt att skatta trädlistor för trakter som planeras att avverkas genom att hämta information från skördardata. Baserat på de tillskrivna trädlistorna kan virkesvärdet på sådana trakter uppskattas antingen med hjälp av apteringssimuleringar, eller enklare, med hjälp av stamprismodeller.

I denna studie testades möjligheten att förbättra den befintliga metoden för utbytesprognoser med hjälp av information om träddiametern. Detaljerad information samlades in med ett mobilt, ryggsäcksmonterat laserskanningssystem (MLS) Systemet integrerar ett tröghetsnavigerings-system, NovAtel SPAN-IGM-S1, och en Velodyne VLP-16 laserskanner för att samla in 3D-punktmolnsdata som ger information om träddiameter och positioner för enskilda träd. Försöksytorna bestod av 5 blandbarrskogsområden på ca 0,16 till 0,32 ha belägna i nordöstra Sverige.

Arbetshypotesen för studien var att informationen som tillhandahölls av MLS-systemet kan användas för att tillskriva trädlistor som bättre matchar utfallet från avverkning, med avseende på diameterfördelningar och därmed förbättra noggrannheten i skattningar av trädvolym och virkesvärde.

Huvudsyftet med studien var alltså att testa arbetshypotesen under flera scenarier, där skattningar av virkesvärdet gjordes med hjälp av olika kombinationer av indata i modellen.

Resultaten tyder på att:

- Matchningen mellan diameterfördelning i skattningar och utfall förbättras avsevärt om man tillför MLS-data som indata i skattningarna
- Osäkerheten i skattning av virkesvärde minskade väsentligt
- Genom att använda dataförstärkning (data augmentation) förbättrades resultaten ytterligare, och minskade avvikelserna mellan skattningar och utfall

Slutsatser:

- Den begränsade utvärderingsmaterialet, fem ytor, gör att det inte går att generalisera resultaten från studien. Men det finns fortfarande bevis för att diameterinformation som hämtats från MLS-mätningar kan ha en positiv effekt på utbytesskattningar, förutsatt att trädslagsfördelningen är känd.
- Förstärkningsmetoden som utvecklats för denna studie har potential att öka skattningsnoggrannheten, speciellt för virkesvärdet, förutsatt att detaljerad rumslig information finns tillgänglig i planeringen före avverkning. MLS är kompatibelt med detta tillvägagångssätt, om det kan insamlas på ett systematiskt sätt.
- Kostnaderna för MLS-undersökningen har inte inkluderats i studien, men vid operativ användning av metoden bör inventeringskostnaderna övervägas för att välja den mest kostnadseffektiva metoden.
- Resultaten öppnar för framtida forskningsriktningar med fokus på:

- Användning av stamkvalitetsvariabler från MLS-data (exempelvis rakhet, gren- och kviststorlek) för skattningar av produktvolymer och utbytesvärde.
- Inkludering av andra typer av hjälpmedel, såsom luftburna laserskanningsdata och / eller satellitbilder. Till exempel mått som beskriver den spatiala fördelningen av skogsvegetationen. Medan satellitbilder kan skatta trädslagsfördelningen som krävs för skattningar av virkesvärde och produktutfall.

1. Project background and research objectives

In a customer-oriented production strategy, wood supply companies must schedule forest cuttings to meet industry requirements. The cutting plans are usually based on yield predictions derived from the information available in the stand databases, and the bucking decisions are made according to price and demand matrices for various timber assortments (Malinen et al 2001, Arlinger et al 2003). Accurate product recovery predictions are essential for efficient planning of forest operations and optimising wood flows in the forest industry (Moberg & Nordmark 2006, Wilhelmsson et al 2011).

When the description of the forest stand is inaccurate, accuracy of the yield estimates will be poor and necessitate additional efforts to comply with the industry requirements (Nordström & Möller 2009). For instance, the information on forest stands available in the forest registers is not spatially explicit and comprises mainly average attributes. Spatially detailed information on tree height, basal area, timber volume, species, stem diameters, and tree height distributions are crucial for accurate value and product recovery predictions.

In Sweden, freely available raster maps containing the most common forest state estimates (i.e., basal area, mean basal area, weighted diameter, and height, standing volume and biomass) are produced by the Swedish University of Agricultural Sciences, using the national airborne laser survey data acquired by the Swedish government agency for mapping and land registration (Lantmäteriet). The map service, called *Skogliga grunddata*, is used by the entire forest sector to improve forestry planning, and serves as a basis for decisions concerning many different and new applications (Anon 2020b). Currently, the forest state attributes provided by *Skogliga grunddata* are also some of the most important inputs for yield predictions in the imputation system developed by Skogforsk (Söderberg 2015, Söderberg et al 2017, 2018).

There is a large body of literature investigating the use of harvester data for product recovery and yield predictions in Nordic countries (Malinen et al 2001, Kivinen et al 2005, Peuhkurinen et al 2008, Holmgren et al 2012, Barth & Holmgren 2013, Sanz et al 2018). In Sweden, Skogforsk has developed a spatial database system where production files from harvesters are uploaded, quality-checked, and compiled in various database tables. In the current technical implementation, the harvester data are also stored by homogeneous, spatially compact microcompartments with an average area around 0.5-1.0 ha (Figure 1). For each microcompartment, the database stores detailed information of all stems and bucked logs, as well as summary statistics such as tree species composition, timber volume, and tree heights. The geographical location of the microcompartments, as well as the harvesters, can also be retrieved, as field information can be easily co-registered to various types of GIS products. The harvester database can therefore provide a large pool of reference observations to support the forest mapping using multisource GIS data.

The yield and product recovery workflow developed by Skogforsk is based on nearest neighbour (NN) imputations, which is a donor-based method for replacing missing data with data with similar characteristics (Eskelson et al 2009). Due to their simplicity and ability to provide operationally useful results, NN imputations have gained strong

acceptance in forestry (Eskelson et al 2009). More precisely, Skogforsk's system uses the k-MSN (Most Similar Neighbour) imputation method (Moeur & Stage 2006) deployed via the 'yaImpute' package (Crookston et al 2007) of the R programming software (R Core Team 2020). Technical details and empirical evaluations of the system are provided by Möller et al (2017), Söderberg et al (2017) and Söderberg et al (2018). This system is currently implemented at several major Swedish forest companies, e.g. Sveaskog and Södra.

Using imputations, tree lists from the harvester database can be assigned to new forest tracts, and the potential yield on such areas can then be estimated using bucking simulations. Bucking simulations based on stem reconstruction harvester data are the best approach for product recovery predictions (Malinen 2007). Alternatively, stem price lists based on species-specific DBH classes and stem volumes can be used for more expeditive value recovery calculations (Möller & Arlinger 2007, Möller et al 2007).

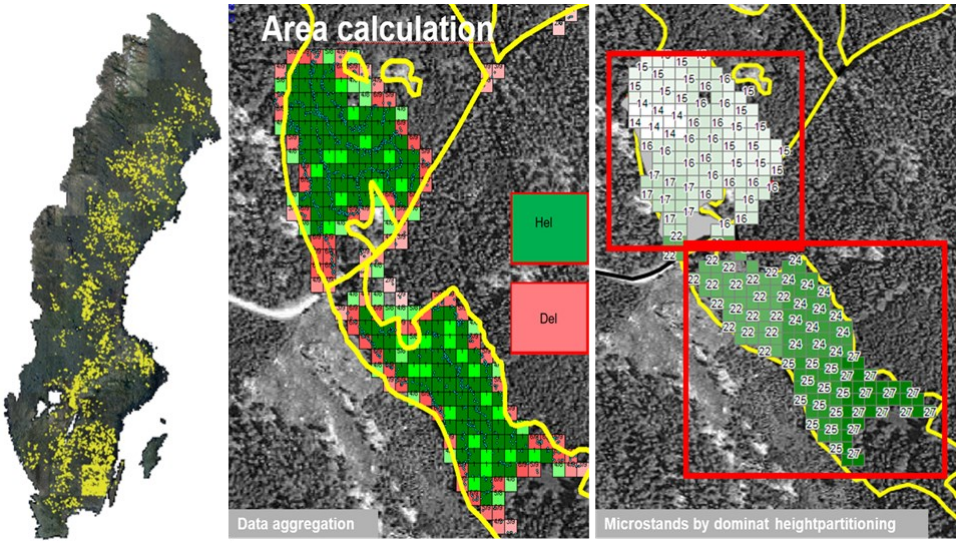


Figure 1. Spatial distribution of the reference observations (yellow dots) currently available in the harvester stem database at Skogforsk (left panel). Forest stand segmentation into microstands based on dominant height measurements provided by the harvester data is illustrated in the centre and right panels. Harvester positions are represented by the black dots and microcompartments with similar properties (tree height in this example) are the numbered rasterized areas shown in different colours. The right panel shows details from harvester data by microcompartments of about 0.5 hectare

For accurate product recovery predictions, it is advantageous if the DBH distributions of the imputed tree lists match the ground-truth DBH distributions as closely as possible. Currently, information on the DBH distributions provided in the forest registers and the forest state estimates (*Skogliga grunddata*) is restricted to average values. Usually, additional information on the entire DBH distributions is coarsely estimated by low intensity pre-harvest inventories.

One of the emerging technologies for automating data acquisition in forest inventories is mobile laser scanning (MLS). MLS are terrestrial measurement systems that produce 3D point cloud data that can be used for algorithmic modelling of individual tree stems (Liang et al 2014). Compared to single viewpoint terrestrial measurement methods for forestry applications, such as terrestrial laser scanning, MLS has the advantage of

continuous data acquisition from multiple viewpoints, reducing the tree occlusion errors, and helps cover wider areas in a cruising mode (Liang et al 2014, Holmgren et al 2019). Our study used a backpack-mounted system integrating a NovAtel SPAN-IGM-S1 INS and a Velodyne VLP-16 scanner for stem positioning and DBH measurements in pre-harvest inventories. The technical set-up and details of the procedure can be found in Holmgren et al (2017) and Holmgren et al (2019).

The research objective was to investigate the use of detailed DBH information provided by MLS technology for improving volume and value recovery predictions, assuming that information on species composition is available. The DBH information can be inserted as auxiliary data into the imputation workflow to search for relevant matches in distribution, not just in the form of averages.

2. Material

The material consists of several datasets:

- tree lists provided by mobile laser scanning (MLS) and from harvesters (HRV)
- auxiliary information in the form of raster maps of predicted forest state attributes derived from airborne laser scanning data
- species-specific stem prices by DBH classes

2.1 Study area

Data was acquired on six experimental blocks of 0.16 to 0.32 ha containing mixed coniferous forest, located in Västerbotten county, northern Sweden. Due to technical issues, the measurements on block no.2 could not be used, so the analyses were run on only five experimental blocks. The broad spatial location of datasets is presented in Figure 2.

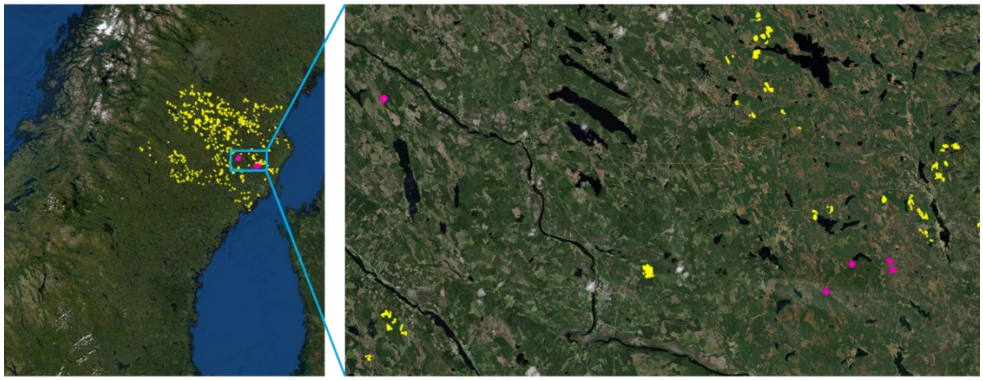


Figure 2. Geographical location of the empirical blocks (Google Earth Pro v.7.3.2, Map data: Google, Image Landsat/Copernicus, 2018). The yellow dots indicate the locations of the training dataset, and the validation data is shown in pink.

2.2 Tree list information

Mobile laser scanning data was collected on selected experimental forest areas (or blocks), while the HRV datasets were available in the Skogforsk's harvester database and for the blocks. The harvester data provides the entire tree list information (e.g., stem volumes, stem dimensions, tree species distribution, and harvester positions), while the MLS contains only the planimetric (x, y) positions for a sample of trees and their DBH. The experimental design and the algorithms for DBH extraction, as well as the accuracy assessments of the results provided by the mobile laser scanning measurements, are presented in Willén et al 2018. Details on the MLS acquisition, data processing, and derivation of DBH measurements are described in Holmgren et al (2017) and Holmgren et al (2019). In the following sections, we assume that the tree-level DBH measurements provided by the mobile laser scanning system and the harvester are error free.

2.2.1 Species-specific stem price lists

Tree values (SEK/m^3 under bark) were calculated using stem price lists for the main tree species (Spruce, Pine and Deciduous) compiled by Skogforsk specialists using industry data from 2020 representative for our study area. The prices for Spruce and Pine trees refer to round wood, while the low-grade wood from coniferous trees and all deciduous trees were aggregated into a common price category (Deciduous). We resorted to this simplification to eliminate the uncertainties related to assigning tree-level quality attributes from the assortment lists in the harvester data to the entire stem.

Table 1. Tree species specific stem price list (SEK/m^3 under bark) by breast height diameter (DBH) classes.

Tree species	DBH class (mm)																					
	80	100	120	140	160	180	200	220	240	260	280	300	320	340	360	380	400	420	440	460	480	500
Spruce	270	270	270	270	331	366	399	410	420	430	435	435	435	435	440	440	445	445	445	445	445	445
Pine	270	270	270	270	310	374	395	411	415	420	425	425	430	430	430	430	430	435	435	423	435	435
Deciduous	250																					

2.2.2 Tree lists on validation blocks

The block-level information from the tree lists provided by each measurement method (MLS and harvester) is summarised in Table 2 and Table 3. The cumulative distributions for the DBH data at block level are presented in Figure 3.

Table 2. Summary statistics – stem number (N), average DBH, standard deviation and the range (min-max) for DBH measurements (in cm) derived from the empirical DBH lists, by block. Block 2 was omitted from the study due to technical issues

Block	Harvester				Mobile laser scanning			
	N	$\overline{DBH}^{(1)}$	$STD^{(2)}$	$Range$	N	$\overline{DBH}^{(1)}$	$STD^{(2)}$	$Range$
Vindeln (Block 1)	457	27.96	7.58	8.00 45.10	102	27.69	8.29	8.10 52.30
Kvarnraningsmyr (Block 3)	580	24.52	8.51	8.60 42.90	163	24.61	7.94	10.00 40.80
Hassjestomyran (Block 4)	356	19.31	6.47	8.10 38.30	114	20.41	7.09	8.40 41.00
Fastkorningen (Block 5)	410	20.44	6.78	8.00 38.10	131	20.52	6.99	8.50 41.50
Furunassjon (Block 6)	293	27.03	6.11	12.80 41.90	81	25.49	6.21	9.50 41.50

(1), (2) *arithmetic mean and standard deviation for the empirical DBH measurements (cm)*

Table 3. Summary statistics for the five blocks in the validation dataset derived from harvester data. The overall average is calculated as area-weighted block averages. Block 2 was omitted from the study due to technical issues

Block ID	Block area (ha)	Block area weigths[1]	Volume by tree species				Block volume
				Spruce	Pine	Deciduous	
Vindeln (Block 1)	0.238	0.231	mc/ha	172.20	51.94	8.28	232.42
			%	74	22	4	100.00
Kvarnraningsmyr (Block 3)	0.316	0.306	mc/ha	39.93	213.00	2.29	255.22
			%	15.6	83.5	0.9	100.00
Hassjestomyran (Block 4)	0.159	0.154	mc/ha	75.25	96.96	11.54	183.75
			%	41	53	6	100.00
Fastkorningen (Block 5)	0.157	0.152	mc/ha	171.78	56.22	13.56	241.57
			%	71	23	6	100.00
Furunassjon (Block 6)	0.163	0.158	mc/ha	184.86	67.23	5.97	258.07
			%	72	26	2	100.00
Overall average [2]			mc/ha	118.72	111.24	7.39	237.35
			%	50.0	47.0	3.0	100.00

[1] *The proportion of individual block areas relative to the total block area*

[2] *The block area- weigthed average*

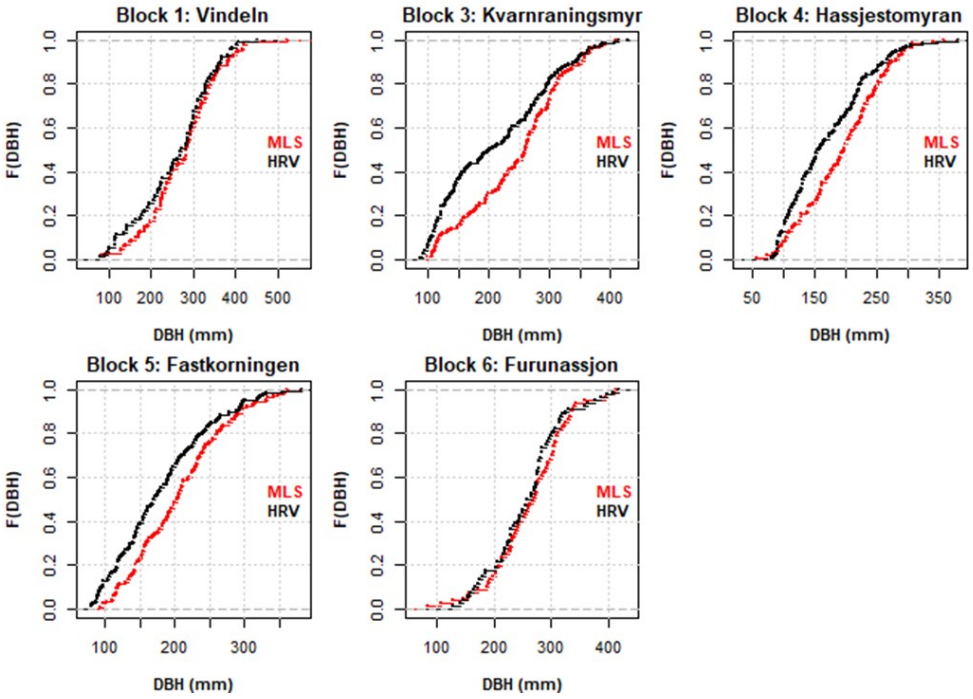


Figure 3. Empirical distribution functions for tree-level DBH measurements produced by the MLS system (red) and by the harvester (black), for each block.

2.2.3 Tree lists in the training dataset

Using Skogforsk's harvester database, 465 harvesting sites were selected as reference data. The stands were segmented into 4737 microstands (yellow dots in Figure 1) using the methodology described by Söderberg et al (2017) and Söderberg et al (2018). Thirty-two percent of the microstands were dominated by Pine and 68% by Spruce. While the empirical blocks had clear spatial delineation in the form of spatial polygons, the boundaries of the observations from the harvester database (microstands or calculation areas) had to be approximated using the harvester positions provided by the GNSS receivers mounted on the machines. For this, 10-m circular buffers were constructed around the machine positions and then merged, and the microstand boundaries were then approximated as the outer polygon surrounding the harvester position buffers. Summary statistics for selected attributes on the tracts and microstands are presented in Table 4.

Table 4. Summary statistics of area and volume for the forest tracts and the microstands in the training dataset.

Aggregation level	Statistics		
	Range [1]	Mean	CV [2]
Area(ha)			
Stands	1.05-41.52	7.85	84.86
Microstands	0.51-2.13	0.77	27.88
Volume (m^3/ha)			
Stands			
Spruce	105.10-489.81	213.56	27.49
Pine	112.22-396.71	188.86	23.27
Overall	105.10-489.91	196.68	26.48
Microstands			
Spruce	105.10-489.81	213.56	28.47
Pine	112.22-396.71	188.86	24.01
Overall	71.97-838.49	258.97	40.19

[1] The minimum and maximum values, with the minimum positive value in the parenthesis

[2] Coefficient of variation relative in percentages relative to the average value

2.3 Auxiliary information

2.3.1 Auxiliary information extracted from the DBH distributions

For the empirical blocks, the DBH distributions are available from the tree lists recorded by the harvesters, as well as from the MLS measurements, while for the database observations, the information on the DBH distribution comes only from harvester data. Preliminary analyses (Holmgren et al 2017) indicate good correspondence between the DBH information provided by the two measurement systems. The DBH information available from the tree lists in the harvester database was considered compatible to MLS measurements and were treated similarly in the analyses. The MLS data were considered as probabilistic samples from the measurements, which is not true, but this assumption is required for justifying the analyses.

The DBH distributions were quantified using various statistics such as rank statistics and cumulants extracted similarly from MLS measurements. The rank statistics consist of four DBH quantiles for the probabilities of 0.106, 0.309, 0.691 and 0.894 corresponding

to the evenly spaced standard normal quantiles of -1.5, -0.5, 0.5, and 1.5. The cumulants (CML) are higher-order statistics derived from the empirical DBH lists as non-linear combinations of the central moments, thereby providing additional information to the usual location, scale, skewness, and kurtosis statistics. Here we used only four cumulants (from the 3rd up to the 6th), defined as:

$$\begin{aligned} CML_3 &= m_3 \\ CML_4 &= m_4 - 3m_2^2 \\ CML_5 &= m_5 - 10m_2m_3 \\ CML_6 &= m_5 - 15m_4 - 10m_3 + 30m_2^3 \end{aligned} \qquad \text{eq. (1)}$$

where $m_k = E\big[dbh - E[dbh]\big]^k$ is the central moment of order k of the empirical DBH distribution and $E[dbh]$ is the expected value. Note that the second central moment of the empirical DBH distribution is biased, since the tree-level DBH observations are not always independent. The lack of independence is due to the various spatial patterns characterising the tree distribution within forest stands. All the higher moments and cumulants, including the second central moment, can be also biased. As far as we know, the effect of spatial autocorrelation on DBH distribution modelling has not been examined in forestry research, and we aim to assess the effect(s) of these biases in this study. A set of eight features describing the DBH data was produced, henceforth denoted as X.MLS.

2.3.2 Auxiliary information extracted from National Forest State Maps

The raster maps for the main forest state attribute estimates were produced between 2009-2014 using the national airborne laser (ALS) survey data acquired by the Swedish government agency for mapping and land registration (Lantmäteriet). The Swedish Forest Agency and the Swedish University of Agricultural Sciences (SLU) combined the laser data and field plots from the Swedish National Forest Inventory to derive digital raster maps of the main attributes characterising the forest state (i.e., basal area, mean basal area weighted diameter and height, standing volume, and biomass) at a spatial resolution of 12.5 x 12.5m. In our study, a set of four features from the National Forest State Map attributes (called here X.SKGD) was used, namely basal area, mean basal area weighted diameter and height, and standing volume. The point estimates for the forest attributes correspond to the 2014 calendar year, and their accuracy is not provided, but a discussion on the lack-of-fit of the models is provided by Nilsson et al. 2017.

2.3.3 Field data configuration

An overview of the spatial datasets for the five blocks used in the analyses is shown in Figure 4. For block 1 (Vindeln), the MLS, harvester and SKGD pixel centre positions do not fully cover the extent of the block, due to errors (most likely recording errors) that have contaminated the field measurements. The field measurements (MLS, HRV and SKGD data) were therefore clipped within a smaller domain. For HRV data, the tree positions are not directly available. Instead, the machine positions are logged and used

for an approximate positioning of the trees within blocks. Based on empirical knowledge, we assume the average tree positioning errors for HRV data should be < 5m. On all plots, the HRV positions could also extend outside the block boundaries (not shown in Figure 4), since the HRV operator had to manoeuvre to access the trees in the blocks.

The areas of three of the five blocks, namely blocks 4, 5 and 6, are substantially smaller (about half the size) compared to blocks 1 and 3. This was not planned, because initially all the blocks were intended to cover approximately equal areas. Unfortunately, due to some unexpected errors affecting the harvester routing on blocks 4,5 and 6, these blocks had to be adjusted to half the original extent.

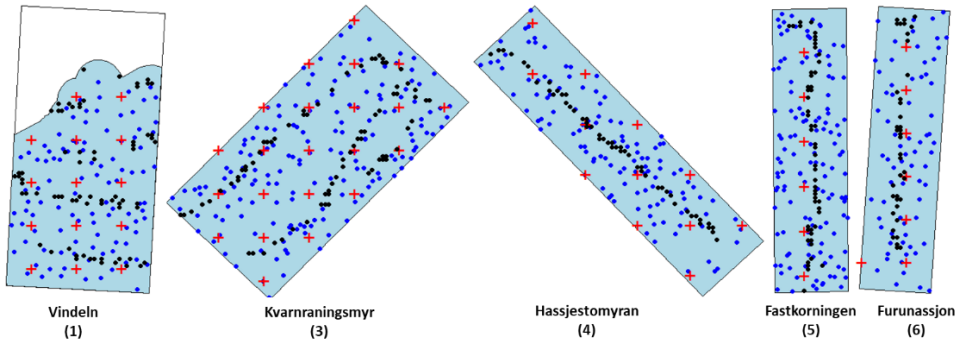


Figure 4. Overview of the block-level spatial datasets. The block boundaries are represented by the rectangular polygons. The tree positions provided by the MLS data are shown with blue dots, the harvester positions with black dots, and the pixel centres for SKGD raster with red crosses.

3. Methods

3.1 Quantisation of DBH distributions and value recovery calculations

MLS and harvester measurements provide datasets containing tree-level records of DBH values. The number of trees can vary substantially between different forest tracts and microstands. In addition, the stem price models (section 2.2.1) were constructed for predefined, discrete DBH classes. In order to obtain comparable DBH distributions and to perform the value recovery calculations, the raw, empirical DBH lists were compressed into a fixed set of discrete DBH classes corresponding to the stem price models in section 2.2.1. This is equivalent to constructing frequency histograms with the number of bins equalling the number of DBH classes. The cuts for defining the histogram bins were selected as the intervals between the DBH classes in the stem price list. The DBH quantisation is not a requirement for value recovery calculations, but it allows a comparison of the reference and imputed tree lists by filtering out the differences in stem numbers.

The species-specific and overall monetary value totals (in SEK) were calculated for microstands and blocks using the generic formulae:

$$\tau_{Vol,c}^{sp} = \sum_{k \in S_c} v_{ck}^{sp} \qquad \tau_P^{sp} = \sum_{c \in \varphi_{DBH}} p_{sp,c} \tau_{Vol,c}^{sp}$$

$$\tau_{Vol}^{sp} = \sum_{c \in \varphi_{DBH}} p_{sp,c}\, \tau_{Vol,c}^{sp} \qquad \tau_P = \sum_{sp \in \varphi_{SP}} \tau_P^{sp}$$

$$\mu_{Vol}^{sp} = \tau_{Vol}^{sp} / Area \qquad \mu_P^{sp} = \tau_P^{sp} / Area$$

$$\tau_{Vol} = \sum_{sp \in \varphi_{SP}} \tau_{Vol}^{sp} \qquad \mu_P = \tau_P / Area \qquad \text{eq.(3}$$

$$\mu_{Vol} = \tau_{Vol} / Area$$

For each diameter class *c*, the species-specific volumes $\tau_{Vol,c}^{sp}$ (m³) were obtained by aggregating the volumes v_{ck}^{sp} for the subsets S_c of trees that belong to certain species *sp*, which were then summed to obtain the total volume by species τ_{Vol}^{sp}. The overall total volumes were obtained by aggregating the species-specific totals. Since the totals are correlated to areas, and the microstand and block have varying areas, the average per-hectare volumes $\mu_{Vol_{sp}}$ and μ_{Vol} were also calculated. The values by species and diameter class were obtained by multiplying the $\tau_{Vol,c}^{sp}$ by the prices $p_{sp,c}$ (SEK/m³) from the price lists. The aggregation and the derivation of the per-hectare average values $\mu_{P_{sp}}$ and μ_P (SEK/ha) follow the same reasoning as for volume calculations.

3.2 Augmenting block-level auxiliary information

The microstands in the training data are generated solely from the tree lists and machine positions in the harvester data. Segmenting larger forest tracts stands into microstands creates more homogeneous areas for post-harvest analyses and reduces the variance of the imputation results, mainly because (1) the correlations between the responses and auxiliary variables tend to increase with the homogeneity of the forest areas, and (2) increasing the size of the training dataset has a positive effect on the accuracy of nearest neighbour imputations.

Using harvester data as input for segmentation is not an appropriate technique for predictive purposes since harvester data is not available for the standing forests. Consequently, only the algorithm for microstand creation can be applied in its current form for the training datasets.

Reasonable questions would be whether or not the field blocks should be further segmented into microstands, and how the segmentation would work in the absence of harvester data. It could be argued that the field blocks have small areas, and a further segmentation into smaller units may not be justified in terms of increasing homogeneity. However, it is known that ensemble methods frequently used in machine learning, such as bagging (bootstrap aggregating), capitalise on averaging overly large sets of predictions produced by predictive models that have relatively low bias but high variance. Averaging such predictions will, theoretically, reduce both the bias and the variance of the aggregated results. Bootstrapping large databases can easily become a computationally demanding task, and the hierarchical structure of the training data (microstands nested into forest) would require specific implementation of the resampling procedure.

However, the idea of increasing the number of predictions at block level for variance reduction does not depend upon bootstrap aggregation. More precisely, if there were a way to produce a large number of predictions within a block, then aggregating the results may increase the accuracy. As in the example of bagging, a main assumption here would be that each of the within-block predictions is approximately unbiased for the block-level

parameter of interest, which is a long shot in the case of near-neighbour imputations (as well as for any other model-based prediction method). Nevertheless, constructing approximately unbiased estimators for the block-level auxiliary averages is technically possible without the need for segmentation into microstands. Using the estimated quantities in the imputations can increase the confidence that the imputed attributes would at least not manifest major biases, provided there is a sufficiently large training dataset. In this case, the variance reduction is guaranteed by aggregation, independently of the amount of bias. We therefore propose a novel strategy for augmenting the block-level predictions using so-called pseudo-plots, which do not rely on harvester information. The approach will be applied to the X.SKGD and X.MLS auxiliaries, in order to increase the number of imputations on each block.

In order to explain the approach, we will first introduce the basic notation and some general concepts related to sampling theory for infinite populations – see Cordy (1993), Stevens & Urquhart (2000) and Mandallaz (2008, §4) for theoretical insights into this topic. We start by considering the simple case of estimating some parameter of a discrete population P (for instance, a population of trees) spread over a planar surface F of area $\lambda(F)$. This would be the typical case for estimation following fixed-area plot sampling in forest inventories. We also assume that the boundaries of F are known (for instance, a polygon in a GIS framework), enabling us to define a sampling frame, and we have access to individual tree positions. The question now is how to select a sample of trees in F for the sake of inference on P, but unlike the discrete case, the inference here has a spatial component. For instance, we would like to infer something about the average volume per hectare, and not about the average tree volume.

We do not intend to provide a full description of the topic here but, since we will be working mostly with per-hectare averages, we will follow the two-step explanation provided by Stevens & Urquhart (2000): (1) create a so-called local density of the attribute of interest at a single sampling point, and (2) use an appropriate estimator to aggregate the local densities obtained for a sample of points selected according to a pre-determined probabilistic sampling design.

An intuitive illustration explaining the local density calculation for an attribute at single random sampling point is described in Figure 5ab. First, note that each element (a red cross) of the discrete population P located within F receives an a priori neighbourhood support (dashed line circles Kr with radius r) that can be interpreted as a geometrical representation of its inclusion density (Figure 5a). The inclusion densities change with the location of the population elements (red crosses) relative to the *F* boundary. In Figure 5b, the local density of the attribute at a sampling point x (the blue dot) is calculated only for the population elements (red crosses) in F located within a neighbourhood (the blue circle) of radius r around x. It is vital that the local densities are design-unbiased (sensú Gregoire & Monkevich 1994) or simply unbiased (Scot & Bechtold 1995), which means that they are constructed at any location in a way that preserves the attribute of interest.

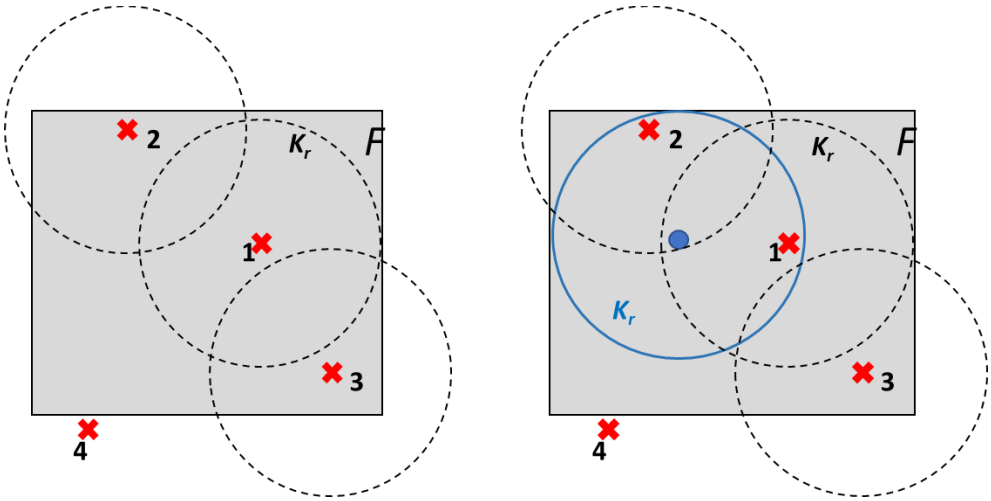


Figure 5. Plot configuration example. Figure in the left depicts four elements of a discrete population (red crosses), three (1,2, and 3) of which are located within a planar surface F (the sampling frame) of area λ(F). The inclusion densities of the points 1,2 and 3 are proportional to the intersection areas between the circles Kr (with radius r) and surface F. Next, a sampling plot (blue circle Kr) located on a random sampling point x (blue dot) containing two population elements is shown in the right display. The local density at the sampling point x is calculated using an aggregation function that balances the inclusion densities of points 1 and 2.

In our particular case, the pseudo-plots sampling approach is applied to two discrete populations, one consisting of all MLS positions and the other of all the pixel centres in the SKGD rasters, across all blocks. The derivation of the local densities is described in Appendix A1.

The partitioning by blocks is equivalent to stratification followed by independent sampling on each block. Basically, this means that the analyses can be run independently by block, and the block-level results can be then linearly aggregated. The same sampling points were used for estimating the local densities of X.MLS and for the X.SKGD auxiliaries. Since the sampling approach relies only on a GIS system, the sample size (i.e., the number of sampling points x) can be set arbitrarily to any reasonable value that is feasible for the user. However, the sample size selection should provide a trade-off between the computational intensity and accuracy (i.e., low sampling variance) and can be easily optimised using appropriate training data. Here, the sample size over all five blocks (approx. 1 ha) was set to 1000 points and adjusted for each block by proportionality to block area.

3.3 The nearest neighbour imputation procedure

The mechanics of nearest neighbour imputations requires two datasets: (1) a reference (or training) dataset that contains two types of variables: the responses (the attributes to be imputed) and feature variables (or auxiliaries), and (2) a target dataset that only consists of auxiliaries. The responses of the target dataset are considered as 'missing' and are selected as being among the responses (or combination of such) in the training dataset. In addition, the auxiliaries in both reference and target datasets must share the same feature space. A thorough discussion on tuning methods for nearest neighbour imputations in forest inventory applications is provided in McRoberts (2009) and McRoberts et al (2015).

For a feature subset Z^q, the similarity measure between a pair of observations (z_i, z_j) is usually expressed using various distance metrics with the generic formulation:

$$d(f_i, f_j, W) = \left[(f_i - f_j)^t W (f_i - f_j)\right]^{\frac{1}{p}} \quad \text{eq. (2)}$$

where $f_{i,i=1:q}$ and $f_{j,j=1:q}$ are feature column vectors, W is a (q×q) weight matrix, and the exponent p is usually chosen as 1 or 2. The similarity metric used for running the imputations was the unweighted Euclidean distance (p=2 and W as identity matrix in eq. 2) on the standardised features. The item-level predictions were obtained by averaging over k=5 neighbours for direct imputations at block-level, and by using k=1 for the pseudo-plots approach. In our case, tuning the imputation by selecting the most appropriate k-values and auxiliaries could be performed by resampling the training data in the harvester database. In the absence of field data for training, the similarity in a high-dimensional feature space cannot be properly evaluated, and alternative tuning procedures must be developed. When using a limited source of auxiliary information (such as X.SKGD), the feature selection step can be skipped, since the feature space dimensionality is limited to four.

At this point, it is worth mentioning that a different approach to imputations based on Most Similar Neighbour (MSN) as a similarity measure was used by Söderberg et al (2017) and Söderberg et al (2018) for yield and product recovery predictions. The MSN imputations rely on canonical correlation analysis that projects the auxiliaries and the responses onto a new coordinate system to construct orthogonal variables with maximal pairwise linear correlations, and the resultant correlation structure is then used as the matrix W in eq. 2. This may complicate the method assessment step, because each feature in eq. 2 would receive a small or large weight when calculating the similarities between training and reference data. For these reasons, the standardised Euclidean distance was adopted instead. The near-neighbour imputations were run using the implementation provided by the 'yaImpute' package (Crookston & Finley 2007) of the R statistical software (R Core Team 2020).

3.4 Analyses and assessment

3.4.1 Scenario analysis

The effect of using MLS information to support near-neighbour imputations was assessed using a sensitivity analysis, where two main scenarios were defined. The first scenario (A) refers to imputations directly on block-level auxiliaries; in the second (B), the imputations using the pseudo-plots approach were grouped under scenarios B1 to B4, corresponding to using pseudo-plots with areas of 150, 250, 500 and 1000-m². The reason for using several pseudo-plot areas was to test the robustness of the approach against the variability in MLS sample size at pseudo-plot level and inclusion densities caused by border effects.

For each of the A and B scenarios, four cases were defined as combinations of different sets of auxiliary information: 1) X.SKGD, 2) X.SKGD + X.MLS.c, 3) X.SKGD + X.MLS.q and 4) X.SKGD + X.MLS.c + X.MLS.q.

3.4.2 Accuracy assessment

The similarity between the ground-truth (DBH) and imputed DBH distributions ($D\hat{B}H$) was assessed using the chi-square histogram distance (Pele & Werman 2010), calculated as $\chi_h^2(H,\hat{H}) = \frac{1}{2}\sum_{i\in h}(H_i - \hat{H}_i)^2/(H_i - \hat{H}_i)$ after quantisation into the same number of bins h corresponding to the DBH classes from the stem price list. This metric has the advantage of reducing the effect of mismatches between large bins (in terms of frequency or number of elements), which are commonly considered as being less important (or influential) than the differences between small bins.

The main criteria for interpreting the imputation accuracy for the attributes of interest were the relative mean deviation (rMD) and relative mean absolute deviation (rMAD), calculated in percentages of the ground-truth attribute as

$\boldsymbol{rMD}(\%) = 100\sum_{i=1:5}\frac{a_i(Z_i-\hat{Z}_i)/Z_i}{\sum_{1=1:5}a_i}$ and $\boldsymbol{rMAD}(\%) = 100\sum_{i=1:5}\frac{a_i|Z_i-\hat{Z}_i|/Z_i}{\sum_{1=1:5}a_i}$. Both rMD and rMAD were calculated as weighted averages of the five blocks, where the weights were proportional with the block areas a. $Z_i$and $\hat{Z}_i$ remain for the ground-truth and the predictions for volume (m³/ha) and monetary value (SEK/ha) on the i^{th}-block, and the term deviations remains for the differences between ground truth and predicted values $(Z_i - \hat{Z}_i)$. Under scenario A (k=5), $\hat{\bar{Z}}$ is the arithmetic mean over the k-neighbours imputed to a block, while under B (where k=1) $\hat{Z}$ is the average over the imputations on the pseudo-plots sampled at block-level. The use of mean absolute deviations instead of root-mean squared deviation on small validation datasets is justified due to its robustness against the influence of large outliers.

4. Results

4.1 DBH distribution predictions

The correspondence between the diameter distributions from the imputed and empirical tree lists (as measured by the harvesters after the final forest cuts), quantified using the chi-square distance, are presented in Table 5.

Table 5. Average chi-square histogram distances between the imputed and ground truth DBH lists after quantisation by the DBH classes used by the stem price lists.

Scenario	Auxiliary information / Cases			
	X.SKGD	X.SKGD X.MLS.c	X.SKGD X.MLS.q	X.SKGD X.MLS.qc
Case:	(1)	(2)	(3)	(4)
A	332.92	86.04	81.48	91.44
B1 (150 m²)	343.73	83.44	83.50	86.01
B2 (250 m²)	343.73	81.87	83.45	86.23
B3 (500 m²)	343.73	86.96	85.13	88.82
B4 (1000 m²)	343.73	89.75	85.85	89.79

On average, the mismatches between the imputed and empirical DBH histograms were systematically smaller when X.MLS auxiliaries (cases 2-4) are included in the feature set. Using the forest state attributes (case 1) produced the largest differences. The best correspondences between the imputed and empirical DBH histograms were obtained under scenarios A/case 3 and B2/case 2 (with bold fonts in Table 5). Using the X.MLS.q (i.e., the DBH quantiles from the MLS measurements, case 3 in all scenarios) seems to provide overall best results. Imputations based on X.SKGD data alone (case 1 in all scenarios), even when controlled for dominant tree species, did not work satisfactorily.

4.2 Volume predictions

Tree species-specific and the block-level relative deviations for volume are presented in Table 6, together with overall rMAE and rME. The largest deviations (approx. -13% and 30% for rMD and rMAD) occurred for scenario A/case 1 (X.SKGD auxiliaries), but including the DBH information (cases 2-4) halved both the rMD (down to 5-6%) and rMAD (13-15%). Using pseudo-plots (scenarios B1-B4) had a minor impact on rMD, which decreased to approx. -11% and 26% (case 1) and varied between -3 and -5% for cases 2-4. rMAD remained stable at 26% for case 1 and varied between 12 and 16% for cases 2-4.

Compared to scenario A, using pseudo-plots decreased rMAD by approx. 10 percentage points (pp) for case 1 in scenarios B1-B4, and approx. 5-20 pp for cases 2 to 4. For rMD, the improvement under scenarios B1-B4 varied between approx. 2-37 pp for cases 2 to 4, and about 15 pp for case 1. The explanation for these relatively low to moderate improvements is that the block areas are quite small, producing low variability in the pixels sampled from the forest state estimates (X.SKGD) by a pseudo-plot, and because most of the new information comes from the X.MLS auxiliaries.

It is also interesting to note that the largest errors under the B-scenarios occurred in blocks 1, 4, and 6, which underwent post-inventory modifications. A plausible explanation may be that the ground-truth information from harvester data (such as tree lists and volumes) on these blocks was contaminated, because separating the trees on sub-block areas is not possible using harvester data alone, since the information on tree positions is only approximated from the harvester locations. This allegation is also supported by the

fact that the DBH distributions provided by HVR and MLS (Figure 3) only match well in blocks 1 and 6, while moderate and large discrepancies occur in blocks 3, 4 and 5 (especially block 3). A very intensive, virtually exhaustive MLS inventory provides a very high measurement accuracy, so an explanation for the discrepancies between the two sets of DBH measurements could be that the tree lists compiled from the harvester data may include trees located outside the blocks and exclude trees that are within the block boundaries.

Table 6. Errors in volume predictions.

Scenario/Block	Auxiliary information / Cases			
	X.SKGD	X.SKGD X.MLS.c	X.SKGD	X.SKGD X.MLS.qc
	(1)	(2)	(3)	(4)
	A) *Direct imputations to block-level auxiliaries*			
1	-49.43	-35.97	-31.67	-31.67
3	-0.46	0.83	0.84	4.15
4	-39.54	-16.60	-21.44	-15.17
5	-43.77	-13.46	-18.46	-25.35
6	-34.58	-0.21	0.09	0.09
Overall rMD	-12.51	-5.34	-5.53	-5.14
Overall rMAD	29.71	13.17	13.67	14.77
	B1) Block-level aggregation: 150 m² circular pseudo-plots			
1	-27.79	-22.66	-28.54	-27.77
3	-14.26	3.25	3.13	3.78
4	-58.69	-23.13	-24.46	-23.67
5	-22.95	-14.34	-16.44	-13.29
6	-15.09	-3.70	-6.57	-4.74
Overall rMD	-10.81	-4.42	-5.44	-4.9
Overall rMAD	25.66	12.54	14.83	13.96
	B2) Block-level aggregation: 250 m² circular pseudo-plots			
1	-27.79	-22.72	-28.74	-26.69
3	-14.26	2.87	2.97	4.21
4	-58.69	-20.43	-19.50	-19.96
5	-22.95	-16.21	-17.97	-14.84
6	-15.09	-0.02	-2.02	-1.43
Overall rMD	-10.81	-4.2	-4.95	-4.31
Overall rMAD	25.66	11.72	13.58	12.99
	B3) Block-level aggregation: 500 m² circular pseudo-plots			
1	-27.79	-22.29	-28.07	-26.16
3	-14.26	3.82	3.95	5.44
4	-58.69	-18.57	-17.53	-18.17
5	-22.95	-12.45	-13.02	-12.06
6	-15.09	3.08	3.45	2.71
Overall rMD	-10.81	-3.46	-3.96	-3.60
Overall rMAD	25.66	11.54	12.89	12.75

	B4) Block-level aggregation: 1000 m² circular pseudo-plots			
1	-27.79	-23.77	-26.42	-25.17
3	-14.26	4.22	3.11	4.38
4	-58.69	-18.28	-17.13	-17.73
5	-22.95	-10.69	-12.25	-13.34
6	-15.09	2.09	2.14	1.56
Overall rMD	-10.81	-3.49	-3.91	-3.77
Overall rMAD	25.66	11.53	11.87	12.14

4.2 Value recovery predictions

The ground-truth stem values on the experimental blocks were evaluated by tree species according to the stem prices in Table 1.

As for volume predictions, a similar pattern can be identified for value recovery predictions (Table 8). The lowest accuracy occurred under scenario A, 24-36% for rMAD and 6-28% for rMD. Once again, the worst results were obtained under case 1 (approx. 36% and 28% for rMAD and rMD), but substantial improvements occurred for cases 2-4, about 23-33 pp for rMAD and 10-65 pp for rMD.

The results under scenarios B1-B4 were quite stable at case level, about 28% and 30% for rMD and rMAD (case 1). Compared to case 1 in scenario A, a slight improvement of approx. 17 pp for rMAD could be observed, while the change in rMD was negligible. Important gains were noticed for cases 2-4, where rMAD decreased by approximately 33-66 pp (to 11-18%), and rMD decreased by 40-67 pp (down to 2-6%).

Across cases, the accuracy criteria under case 3 (DBH quantiles) were systematically better under scenarios B1-B4, followed closely by the combination of DBH cumulants and quantiles X.MLS.qc (case 4). Using only the DBH cumulants (case 3) produces best results under scenario A.

Overall, using pseudo-plots with areas of 250 m² produced a good balance between rMAD and rMD. Too small pseudo-plots will sample only a few MLS measurements, making the cumulants and quantile estimates not very reliable. Larger pseudo-plots will tend to over-smooth the auxiliaries, as well as amplify the boundary effects.

Table 8. Species-specific and aggregated value recovery prediction.

Scenario	Auxiliary information / Cases			
	X.SKGD (1)	X.SKGD X.MLS.c (2)	X.SKGD X.MLS.q (3)	X.SKGD X.MLS.qc (4)
A) *Direct imputations to block-level auxiliaries*				
Overall rMD	*27.94*	*7.62*	*9.61*	*6.49*
Overall rMAD	*35.84*	*27.50*	*27.03*	*23.53*
Spruce	31.25	15.60	15.50	13.66
Pine	-3.95	-9.94	-9.71	-8.52
Deciduous sp.	0.64	1.96	3.82	1.35
B1) Block-level aggregation: 150 m² circular pseudo-plots				
Overall rMD	*28.02*	*5.78*	*4.13*	*5.03*
Overall rMAD	*29.71*	*13.85*	*10.86*	*11.17*
Spruce	28.23	8.88	6.71	7.36
Pine	-0.85	-4.04	-3.36	-3.07
Deciduous sp.	0.63	0.94	0.79	0.74
B2) Block-level aggregation: 250 m² circular pseudo-plots				
Overall rMD	*28.02*	*5.46*	*3.43*	*3.68*
Overall rMAD	*29.71*	*13.93*	*10.45*	*10.59*
Spruce	28.23	8.84	6.24	6.42
Pine	-0.85	-4.23	-3.51	-3.46
Deciduous sp.	0.63	0.86	0.70	0.71
B3) Block-level aggregation: 500 m² circular pseudo-plots				
Overall rMD	*28.02*	*3.18*	*2.11*	*2.78*
Overall rMAD	*29.71*	*18.40*	*14.57*	*16.00*
Spruce	28.23	9.18	7.70	7.80
Pine	-0.85	-6.64	-6.23	-5.63
Deciduous sp.	0.63	0.64	0.64	0.61
B4) Block-level aggregation: 1000 m² circular pseudo-plots				
Overall rMD	*28.02*	*3.69*	*3.26*	*3.67*
Overall rMAD	*29.71*	*16.75*	*16.50*	*14.62*
Spruce	28.23	9.64	9.07	8.44
Pine	-0.85	-6.53	-6.62	-5.47
Deciduous sp.	0.63	0.59	0.81	0.70

5. Discussion and conclusions

The main objective of the study was to learn more about the use of mobile laser scanning data for volume and value recovery based on nearest neighbour imputation. The MLS data were limited to DBH measurements and tree positions, and they were not identical to the harvester data obtained on the same areas. At best, the tree lists provided by MLS could be considered as being produced by a highly intensive pre-harvest inventory following a probabilistic design. In fact, the probability sampling is an assumption only required to upscale the sample-level information.

Overall, the DBH auxiliaries seem to have the dominating effect on accuracy. Using the X.MLS auxiliaries in addition to the forest state estimates in X.SKGD will generate more relevant near-neighbours among the reference observations, which will also match distribution (for DBH) and not only in the averages. In addition, X.MLS contains information that directly relates to the trees, and not proxies as X.SKGD. For value recovery, using the DBH auxiliaries with the pseudo-plots approach produces higher gains compared to volume predictions. Arguably this result depends strongly on the specific price lists used for value recovery, as well as on uncertainties in the species composition data. From this perspective, our results may be quite optimistic, since the stem prices we used were rather flat, and we controlled for tree species composition.

Both types of auxiliary information extracted from DBH data (i.e., quantile and cumulants) produced nearly identical results. Using block-level imputations (scenario A in our study), using both the cumulants and the quantiles, seems to be a good option. With the pseudo-plots approach, using the cumulants may be more computationally advantageous and eliminates selection of the quantiles from the tuning process, an objective method for scaling up the cumulants still needs to be found. For this reason, using *a priori* selected quantiles may be a safer option in general.

MLS data seem to underrepresent the small DBH categories. We do not have enough information to conclude whether this is a method limitation or a result specific to our study, but it is reasonable to assume that future technological developments will surely eliminate such issues. Nevertheless, it is unlikely that MLS will manage to provide a full census of the trees in a forest, a more realistic scenario being that the MLS data acquisition will be performed in a sampling approach. Probabilistic sampling methods should be sought, because they are objective and provide support for proper inference.

Another relevant finding of the study is related to the data augmentation idea based on pseudo-plots. Even for our limited validation dataset, this approach looks promising. The pseudo-plots approach has several advantages – it does not require the development of a segmentation algorithm for generating microstands, it relies on a sound statistical framework, and can be computationally efficient for large datasets due to the ease of parallelisation. Further research should investigate methods for selecting the optimal pseudo-plot area, and a cross-validation approach would be a good start, assuming that the same type of spatial information is available both the training and validation data. For instance, creating small-area pseudo-plots on the microstands in the training data would be possible only if the tree positions are available in the harvester data, otherwise such pseudo-plots should cover a sufficiently large area to absorb the positional errors of the trees.

The costs of MLS data acquisition were not considered in our analyses, but the inventory costs may be justified by the substantial value recovery benefits resulted from injecting the DBH information. However, MLS is currently a rather complicated solution when the sole aim is to acquire DBH measurements. Accurate estimates of the true DBH distributions at forest tract stand level can be obtained using much simpler, low-cost sampling methods. A big strength of the MSL measurements is the ability to capture proxy information for stem quality, which otherwise could require more laborious field measurements (Murphy et al 2020). This information could be then inserted into the imputation workflow, in a similar way as with the DBH data, to further refine the value recovery calculations and for product recovery predictions using bucking simulations. We could not perform such analyses in this study because the quality information data was not available. This a major limitation of the study, in addition to the small and contaminated validation dataset.

The imputations benefited from using error-free species information for stratifying the training and validation data during imputations. The stratification decision was justified due to the lack of register data usually used for this purpose. In practical applications, a perfect stratification, even only by the dominant tree species, is surely not possible since the information in the forest registers and from any cartographic product (such as vegetation maps) contains errors. Using the tree species information is critical for value and product recovery, since the prices and the specifications for wood products vary not only by dimension and quality, but also by tree species. When used with uncertain tree species information, the imputation accuracy will likely worsen. These aspects could not be properly addressed here, mainly due to the limited validation data, and further studies are needed.

The study also assumed that the DBH measurements provided by the MLS and HRV are error-free. This assumption may hold for harvester data, since the calibrated harvester measurements are usually very accurate, but the expected DBH measurement errors from MLS would be in the range of about 10 mm (Willén et al 2018). A realistic measurement error propagation into the imputation framework could possibly be performed using Monte-Carlo simulations, provided that a model relating the measurement error to the tree size is available in the future.

In conclusion, using pre-harvesting inventory data (such as MLS or any other sampling technique) may be useful for increasing the accuracy of near-neighbour imputations for value recovery and for forest attribute predictions, assuming that it is possible to control for the tree species compositions. Also, if the sampling design of the pre-harvesting inventory allows, it would be more advantageous to rely on imputations based on the proposed augmentation method to increase accuracy.

References

Anon 2020*a*. Lantmäteriet. GSD-Höjddata, grid 2+. (URL: https://www.lantmateriet.se/sv/Kartor-och-geografisk-information/Hojddata/GSD-Hojddata-grid-2/ (last accessed on 14 December 2020).

Anon 2020*b*. Skogsstyrelsen. Available online at https://www.skogsstyrelsen.se/skogligagrunddata (last accessed on 14 December 2020).

Arlinger J, Moberg L & Wilhelmsson L 2003. Predictions of wood properties using bucking simulations software for harvesters. In: Nevpeu G (ed) Proceedings from the fourth meeting IOFRO Wp5.01-04 "Connections between forest resources and wood quality: modelling approaches and simulation software", Fourth Workshop, Hot Springs 2002. INRA, Nancy.

Barth A & Holmgren J 2013. Stem taper estimates based on airborne laser scanning and CTL harvester measurements for pre-harvest planning. International Journal of Forest Engineering, 24, 161-169.

Cordy CB 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Statistics & Probability Letters 18, 353-362.

Crookston NL & Finley AO 2007. yaImpute: An R Package for k-NN Imputation. Journal of Statistical Software 23,1-16.

Eskelson BNI, Temesgen H, Lemay V, Barrett TM, Crookston NL & Hudak AT (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research, 24, 235-24

Gregoire TG & Monkevich NS 1994. The reflection method of line intercept sampling to eliminate boundary bias. Environmental and Ecological Statistics, 1,219-226.

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

Holmgren J, Tulldahl HM, Nordlöf J, Nyström M, Olofsson K, Rydell J & Willen E 2017. Estimation of tree position and stem diameter using simultaneous localization and mapping with data from a backpack-mounted laser scanner. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-3/W3, 2017 Frontiers in Spectral Imaging and 3D Technologies for Geospatial Solutions, 25–27 October 2017, Jyväskylä, Finland.

Holmgren J, Tulldahl M, Nordlöf J, Willen E & Olsson H 2019. Mobile Laser Scanning for Estimating Tree Stem Diameter Using Segmentation and Tree Spine Calibration. Remote Sensing, 11, 2781 (https://doi.org/10.3390/rs11232781).

Kivinen V-P, Uusitalo J & Nummi T 2005. Comparison of four measures designed for assessing the fit between the demand and output distributions of logs. Canadian Journal of Forest Research, 35, 693–702 (https://doi.org/10.1139/x04-196).

Malinen J, Maltamo M & Harstela P 2001. Application of Most Similar Neighbor inference for estimating marked stand characteristics using harvester and inventory generated stem databases. International Journal of Forest Engineering, 12, 33-41 (https://doi.org/10.1080/14942119.2001.10702444).

Malinen J, Kilpeläinen H, Piira T, Redsven V, Wall T & Nuutinen T 2007. Comparing model-based approaches with bucking simulation-based approach in the prediction of timber assortment recovery. Forestry: An International Journal of Forest Research, 80, 309-321, https://doi.org/10.1093/forestry/cpm012.

McRoberts RE 2009. Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sensing of Environment, 113, 489-499.

McRoberts RE, Næsset E & Gobakken T 2015. Optimizing the k-Nearest Neighbors technique for estimating forest aboveground biomass using airborne laser scanning data. Remote Sensing of Environment, 163, 13-22.

Moberg L & Nordmark U 2006. Predicting lumber volume and grade recovery for Scots pine stems using tree models and sawmill conversion simulation. Forest Products Journal, 56, 68-74.

Moeur M & Stage AR 1995. Most Similar Neighbor: An improved sampling inference procedure for natural resources planning. Forest Science, 41, 337-359.

Möller, J.J., Arlinger, J., Wilhelmsson, L., Sondell, J. & Moberg, L. 2007. Modell för automatisk kvalitetsbedömning vid virkesmätning med skördare. [Model of automatic quality determination for grading using harvesters.] Skogforsk, Uppsala, Arbetsrapport No 642, 2007, 14 pp. (In Swedish.).

Möller, J.J. & Arlinger J. 2007. Praktisk test av automatisk kvalitetssättning vid betalningsgrundande skördarmätning hos Södra skogsägarna i Götaland och Sveaskog i Bergslagen. No 643, 2007, 44 pp. (In Swedish.)

Möller J, Arlinger J, Bhuiyan N, Eriksson I & Söderberg J 2017. Forecasting of log product yield based on forest and harvester data – A description of models and system for creating stem-files and imputation of product yield. Work Report 961-2017, Skogforsk (https://www.skogforsk.se/contentassets/58f58eabf1a44c76822ab66bedaed23b/arbetsrapport-961-2017.pdf, in Swedish).

Murphy GE, Acuna MA & Dumbrell I 2010. Tree value and log product yield determination in radiata pine (Pinus radiata) plantations in Australia: comparisons of terrestrial laser scanning with a forest inventory system and manual measurements. Canadian Journal of Forest Research, 40, 2223-2233.

Næsset E 2014. Area-based inventory in Norway – from innovation to an operational reality. In Maltamo M, Næsset E, Vauhkonen J (eds). Forestry Applications of Airborne Laser Scanning. Springer Netherlands, Dortrecht, 215-240.

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 Sensing of Environment 194, 447-454.

Nordström M & Möller JJ 2009. Den skogliga digitala kedjan - fas 1. Technical report 676-2009, Skogforsk (in Swedish).

Pele O & Werman M 2010. The Quadratic-Chi Histogram Distance Family. In K. Daniilidis, P. Maragos, N. Paragios, Eds.: Berlin, Heidelberg: ECCV 2010, Part II, LNCS 6312, 2010, pp. 749-762.

Peuhkurinen J, Maltamo M & Malinen J 2008. Estimating species-specific diameter distributions and saw log recoveries of boreal forests from airborne laser scanning data and aerial photographs: a distribution-based approach. Silva Fennica 42, 625-641.

R Core Team 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL (https://www.R-project.org/).

Sanz B, Malinen J, Leppänen V, Valbuena R, Kauranne T & Tokola T 2018. Valuation of growing stock using multisource GIS data, a stem quality database, and bucking simulation. Canadian Journal of Forest Research, 48, 888-897 (dx.doi.org/10.1139/cjfr-2017-0172).

Särndal C-E, Swensson B & Wretman J 2007. Model-assisted survey sampling (2nd Ed). Springer, New-York, 710p.

Söderberg J 2015. A method for using harvester data in airborne laser prediction of forest variables in mature coniferous stands. Unpublished Master's thesis, Swedish University of Agricultural Studies (SLU), Umeå, Sweden, ISSN 1401-1204 (https://stud.epsilon.slu.se/8496/7/soderberg_%20j_150918.pdf).

Söderberg J, Willén E, Möller J, Arlinger J & Bhuiyan N 2017. Evaluation of yield forecasts produced by forest laser estimations and harvester data – results from three case studies. Work Report 937-2017, Skogforsk (URL: https://www.skogforsk.se/contentassets/ed3d076f264040fca6834692aeeba420/utvardering-av-utbytesprognoser-med-skogliga-laserskattningar-och-skordardata-arbetsrapport-937-2017.pdf, in Swedish)

Söderberg J, Möller J & Willén E 2018. Evaluation of yield prediction with harvester data. Work Report 981-2018, Skogforsk, (https://www.skogforsk.se/cd_48e5fc/contentassets/4a34a91c429f46b09a9f3abc5fcbc242/arbetsrapport-981-2018.pdf)

Scott, CT and Bechtold WA 1995. Techniques and computations for mapping plot clusters that straddle stand boundaries. Forest Science Monographs 31, 46-61.

Stevens DL & Urquhart NS 2000. Response designs and support regions in sampling continuous domains. Environmetrics, 11, 13-41.

White JC, Wulder MA, Varhola A, Vastaranta M, Coops NC, Cook BD, Pitt D & Woods M 2013. A best practice guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. Natural Resources Canada, Canadian Forest Service, Canadian Wood Fibre Centre, Victoria, BC. Information Report FI-X-010. Available online https://cfs.nrcan.gc.ca/publications/download-pdf/34887 (last accessed on 14 January 2019).

Wilhelmsson L, Arlinger J, Hannrup B, Norström M, Øvrum A, Gjerdrum P 2011. Deliverable D3.5 - Methods and models for relating properties and storage conditions to process efficiency and product quality. Technical Report 750-2011, Skogforsk (in Swedish). In *Intelligent distributed process utilization and blazing environmental key* project (Indisputable Key project no 34732) co-funded by the European Commission within the Sixth Framework Programme (2002-2006).

Willén, E., Söderberg, J., Holmgren, J., Tulldahl, M., Nordlöf, J., Öhgren, J., Rydell, J. 2018. Demonstrations of mobile laser scanning for tree mapping. Report 992. Skogforsk

Liang X, Kukko A, Kaartinen H, Hyyppä J, Yu X, Jaakkola A & Wang Y 2014. Possibilities of a Personal Laser Scanning System for Forest Mapping and Ecosystem Services. Sensors 2014, 14, 1228-1248.

Appendix A1. Local density calculations for the data augmentation procedure

The local densities at a position x are obtained as:

$$\bar{Y}_{K_r}(x) = \sum_{t_i \in K_r} y(t_i) g(x, t_i),$$

where $y(t_i)$ is the attribute of interest y for population element located at position t_i, K_r is the neighbourhood support, and $g(x, t_i)$ is the aggregation function that depends not only on the sampling point location x, but also on the population element position (i.e., the tree position in F). The aggregation function g returns the inverse of the proportion of overlapping area between the neighbourhood support K_r and F, defined as:

$$\pi(t_i) = \lambda(K_r \cap F)/\lambda(F).$$

Therefore, $g(x, t_i)$ can be written as

$$g(x, t_i) = 1/\pi(t_i),$$

and the local density at x becomes a weighted average calculated as

$$\bar{Y}(x) = \left[\sum_{t_i \in K_r} y(t_i)/\pi(t_i)\right] / \left[\sum_{t_i \in K_r} 1/\pi(t_i)\right].$$

The process can be then repeated for a sample S of points (i.e., many blue dots in Figure 5) selected using uniform random sampling on F. Under this simple, probabilistic sampling design, an approximately unbiased estimator for the average local density of the population attribute $\bar{Y}$ on F is then obtained as a ratio of two estimated totals

$$\hat{\bar{Y}} = \left[\sum_{k \in S} \bar{Y}(x)/\pi(x_k)\right] / \left[\sum_{k \in S_k} 1/\pi(x_k)\right],$$

where the nominator is the Horvitz-Thompson estimator for spatial continuum (Cordy 1993), and the denominator is an estimate of $\lambda(F)$.

For X.SKGD, the calculations follow directly the example above. Regarding the MLS data, each pseudo-plot provided an estimate for the DBH quantiles at block level, as well as a set of cumulants. The quantiles estimators $\widehat{M}$ were calculated as in Särndal et al (2003, §5.11), for the ordered set of DBH values $y_{1:s} < y_{2:s} < \cdots < y_{N:s}$ selected at a sampling point x using the cumulative sum $B_l = \sum_{j=1}^{l} 1/\pi(t_{j:s})$ as:

$$\widehat{M}(x) = \begin{cases} y_{l:s}, & if\ B_{l-1} < q\widehat{N} < B_l \\ 0.5(y_{l:s} + y_{l:s}), & if\ B_l = q\widehat{N} \end{cases} \qquad \text{eq.(A1)}$$

where q is a selected quantile and $\widehat{N} = \sum_{t_i \in K_r} 1/\pi(t_i)$ is the estimated number of trees at sampling point x. For the cumulants, which are non-linear statistics, we do not have an unbiased estimator as the one in equation A1. Therefore, they were calculated using equal weighting (i.e., $\pi = 1$) of the DBH values selected at each sampling point.