ARBETSRAPPORT 1075-2021

Tree species mapping for value recovery predictions combining harvester and remotely sensed data

Trädslagskartläggning för utbytesprognoser med hjälp av skördar- och fjärranalysdata

FOTO: ERIK VIKLUND

skogforsk

Liviu T. Ene & Jon Söderberg

Contens

skogforsk

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

Kvalitetsgranskning (Intern peer review) har genomförts den 9 mars 2021 av Erik Willén, Processledare. Därefter har Magnus Thor, Forskningschef granskat och godkänt publikationen för publicering den 18 mars 2021.

 ISSN 1404-305X

Preface

This project developed methods for utilising harvester data as *in-situ* data when tree species are mapped using remote sensing data, satellite data and LiDAR. Improved tree species mapping is crucial in view of its significant impact in yield predictions prior to forest operations to meet industry requirements. The project duration was 2019-2020. The Nils and Dorthi Troedsson research foundation contributed with funding for the project. We also wish to thank SCA Forest AB for the use of harvester data in the project.

Uppsala, March 2021

Liviu T. Ene & Jon Söderberg

Sammanfattning

Målet med projektet var:

(1) att införa en innovation genom att utveckla en robust metod för kartläggning av trädslag i operativt skogsbruk, och

(2) att bedöma om en mer exakt information om trädslag skulle förbättra utbytesprognosens noggrannhet.

Vi har utvecklat ett arbetsflöde för att automatisera de viktigaste databehandlingsstegen som kombinerar *in-situ* skördardata och GIS-produkter som multitemporala Sentinel-2-satellitdata, Skogliga grunddata och trädhöjdsraster från laserdata.

Volymbaserade trädslagsandelar beräknas för tall, gran och löv. Dessutom utvecklades en beräkningseffektiv rutin för att kombinera multitemporal datamängder för att minimera dataförlusterna på grund av moln i satellitdata.

Värdet av trädslagsinformation för utbytesprognoser bedömdes i tre scenarier:

(A) baslinjefall utan trädslagsinformation,

(B) att använda skattade trädslagsandelar, och

(C) använda information om "perfekt" trädslag (avverkade trädslag från skördardata).

Trädslagsinformation användes som hjälpdata i imputeringsrutinerna för utbytesprognoser, antingen som proportioner av stående volym per hektar eller som indikatordata (0-1) för det dominerande trädslaget. Avkastningsberäkningarna framställdes med hjälp av stamprislistor för de viktigaste trädslagen sammanställdes av Skogforsk med hjälp av branschdata från 2020 som var giltiga för vårt studieområde.

Det föreslagna arbetsflödet är robust, bygger uteslutande på programvara med öppen källkod och kan enkelt skalas upp för att hantera större studieområden. Dessutom möjliggör det en bättre användning av *in-situ* data i skogsområdena täckta av moln. Resultaten visar ca. 100% minskning av de absoluta avkastningsförlusterna för gran och tall och ca. 20% minskning för lövträd när den förutsagda trädslagsinformationen införlivades i imputationerna. Avvikelsen från det ideala fallet med felfri information om trädslag och volym reducerades med upp till 62% för gran, 59 % för tall och med 29 % för löv när de skattade trädslagsandelarna användes i utbytesprognoser.

Sammantaget visar resultaten att tillförlitlig information om trädslagsandelar är avgörande för att öka noggrannheten för avkastningsprognos. När det gäller effekterna för skogsindustrin bör resultaten ses som en konservativ skattning utbytesprognoser eftersom mervärdet av trädslagsinformation förväntas öka längre fram i virkesförsörjningskedjan.

Summary

The goal of the project was twofold;

(1) to introduce a major innovation by developing a robust and accurate methodology for tree species mapping in operational forestry, and

(2) to assess whether more accurate information on tree species would improve the accuracy of yield prediction.

We successfully developed a workflow for automating the main data processing steps that combines in-situ harvester data and GIS products such as multi-temporal Sentinel-2 imagery, forest state estimates and height vegetation maps. Specific methods were employed for predicting the tree species proportions (relative to total standing volume). The three main groups of species considered were Spruce, Pine and Deciduous, which are of special interest for the wood industry. In addition, a computational-efficient routine was developed for combining multi-temporal datasets for minimising data losses caused by cloud occlusions in satellite imagery.

The **value of tree species information on yield predictions** was assessed using three cenarios;

(A) baseline case with no species information;

(B) using tree species predictions, and

(C) using 'perfect' information on tree species (the 'ground-truth' data).

Species information was incorporated as auxiliary data in the imputation routines, either as proportions of standing volume per hectare, or as indicator data (0-1) for the dominant species. The yield calculations were performed using stem price lists for the main tree species compiled by Skogforsk specialists, using industry data from 2020 that applied to our study area.

The proposed workflow is robust, relies exclusively on open-source software, and can be easily scaled-up for handling larger study areas. The workflow also enables better use of 'ground-truth' data in the forest areas covered by clouds. The results show virtually a 100% reduction of absolute yield losses for Spruce and Pine and an approximately 20% reduction for Deciduous species when the predicted tree species information was incorporated into the imputations. The gap between the baseline and ideal scenario C was reduced by up to 62% for Spruce, 59% for Pine and 29% for Deciduous when incorporating the predicted species-specific volumes into the imputation routines.

Overall, the findings demonstrate that reliable information on species proportion is vital for increasing the accuracy of yield prediction. With regard to the effects for the forest industry, the results should be seen as a lower bound, since the added value of tree species information is expected to compound further down in the wood supply chain.

1. Introduction

In recent years, Skogforsk has been actively supporting the digitalisation process in the Swedish forest industry. An important outcome of this process is the increased use of harvester data for feedback during forest operations, but also to improve yield predictions based on historical data. Yield predictions are essential for the wood supply companies, enabling them to schedule forest cuttings in a way that matches industry requirements. These cutting plans are usually based on yield predictions derived from information available in the stand databases, and the bucking decisions for cut-to-length logging are made in accordance with price and demand matrices for various timber assortments.

The main challenge today is the quality of the information available in the stand databases. When the description of the forest stand is inaccurate, the yield estimate will be poor, and extra work will be needed to comply with industry requirements. Detailed information on tree height, basal area, timber volume and stem diameter is crucial for accurate yield predictions. These attributes may be efficiently and accurately predicted at stand level (even better than most available forest stand data) by using airborne laser scanning (ALS) data, such as *Skogliga grunddata* provided by the Swedish Forest Agency. Equally important is the prediction of tree species composition, but detailed and highly accurate tree species mapping is currently lacking, potentially causing significant yield prediction errors.

The project idea was to develop a novel system for tree species mapping, combining freely available multitemporal, multispectral satellite imagery and ground-truth information provided by harvester data. This would make it possible for forest companies to improve their mapping of tree species in a very cost-efficient way, and significantly improve their yield predictions. The mapping can also be useful in forest management and planning as well as for mapping of “green infrastructure” and other environmental considerations.

The major innovation of the project is the use of a database containing large amounts of high quality *in-situ* forest data to support forest mapping. Data is collected with calibrated harvester measurements superior to any other *in-situ* measurements in terms of accuracy of measurements and number of observations. As the harvester data is collected continuously, this approach can provide a long-term solution for detailed forest mapping in combination with multitemporal, multisource GIS data. The main result of the project is a detailed procedure for tree species mapping with an accuracy that satisfies the requirements for forest planning operations and optimisation of the wood flow to industry. Also, the value of tree species information had been proved to be paramount for accurate yield predictions based on nearest-neighbour imputations.

The expected impact is a more accurate and dynamic industry supply of timber and pulpwood with the requested species specifications. That would make it possible to improve the optimisation of the supply chain from forest to industry on a tactical planning horizon rather than the more reactive planning that is often the case today due to lack of precision in yield forecasts.

2. Material

The material contains two main types of data: (1) ground-truth information at tree-level obtained from the harvester production files, and (2) auxiliary information comprising multisource GIS datasets in raster format such as Sentinel-2 satellite imagery and various cartographic products airborne laser scanning data. The datasets were acquired for a study area of 53,384 ha across Västernorrland and Jämtland regions, an area managed by *Svenska Cellulosa Aktiebolaget* (SCA).

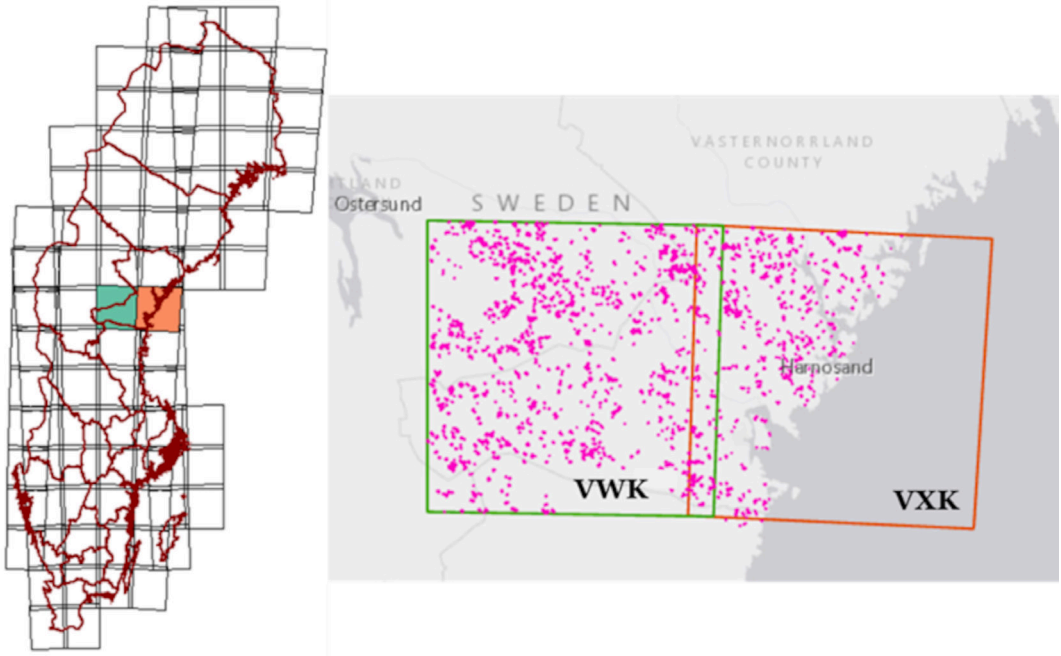

Figure 1. Sentinel-2 scenes outline across Sweden (left panel), with the selected regions of interest in green (scene 33VWK) and orange (scene 33VXK). The right panel contains an overview of the spatial location of the harvested forest stands within the study area.

2.1 HARVESTER DATA

Harvester data from 345 forest stands that had been clear cut prior to Sentinel-2 and ALS data acquisitions were obtained, and then segmented into 2088 microstands (Figure 2) using the methodology described by Söderberg et al. (2017) and Söderberg et al. (2018). Tree species proportions in standing volume were calculated at stand and microstand level. The distribution by dominant tree species is summarised in Table 1:

Table 1. Number of observations (stands and microstands) by the dominant tree species. The values in parentheses represent percentages relative to totals.

Sample size	Spruce	Pine	Deciduous	Contorta	Toal
Stands	245 (71.01)	81 (23.5)	3 (0.87)	16 (4.64)	345
Microstands	1316 (63.03)	624 (29.89)	7 (0.34)	141 (6.76)	2088

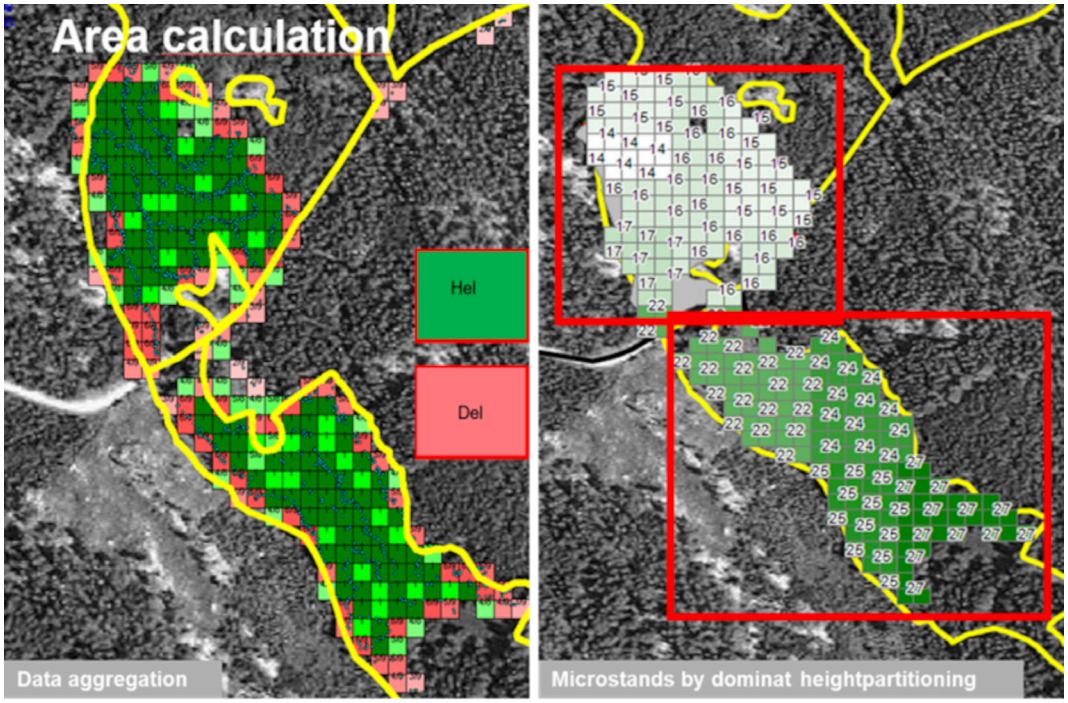

Figure 2. Figure.2 Forest stand segmentation into microstands based on dominant height provided by harvester data. The microstands are derived by aggregating 13x13 m polygons based on their type of spatial connectivity ('HEL' in green color and 'DEL' in red color)..

In the analyses, microstands dominated by Scots pine and lodgepole pine, respectively, were merged into a single Pine class, because the price models (see section 2.2) for lodgepole pine were not available. A summary is shown in Table 2.

Table 2. Area (ha) and volume (m^3/ha) distributions for forest stands and microstands.

Attribute		**Statistic**		
		Range (1)	**Mean**	**CV** (2)
Area (ha)				
Stands		0.69-87.90	6.70	131.44
Microstands		0.69 -3.31	1.11	29.98
Volume (m^3/ha)				
Stands				
	Spruce	0(0.99) -500.29	136.56	59.23
	Pine	0(0.23) -293.71	72.35	88.23
	Deciduous	0(0.01) -128.08	19.39	107.63
	Overall	50.57-1001.56	345.47	38.63
Microstands				
	Spruce	0(0.07) -745.86	147.78	72.69
	Pine	0(0.05) -574.91	93.47	95.54
	Deciduous	0(0.02) -196.98	17.71	125.02
	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

2.2 STEM PRICE MODELS

The yield calculations were performed using stem price lists for the main tree species compiled by Skogforsk specialists, using industry data from 2020 applicable to our study area. The prices for spruce and pine refer to round wood, while the low-grade wood from coniferous trees and all deciduous trees was aggregated into a common price category. 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 3. 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	250	250	250	250	250	250	250	250	250	250	250	250	250	250	250	250	250	250	250	250	250

2.3 AIRBORNE LASER SCANNING DATA

The national ALS data survey *Laserdata Skog* is carried out by Lantmäteriet, with an updating frequency of about seven years. A new data acquisition campaign started in 2018 and will cover 75% of the Swedish area (approx. 350,000 km^2) with 1-2 points per m^2 (Anon 2020a). The first ALS returns are processed by the Swedish Forest Agency to obtain a canopy height model (CHM) in the form of a 2-m spatial resolution raster (*"Trädhöydsraster Laser"*), thereby providing a 3D description of the forest canopy surface (Anon 2020b).

2.4 SATELLITE IMAGERY DATA

Satellite data related to our study area correspond to two Sentinel-2 tiles 33VWK and 33VXK (Figure 1). The sensor has a swath of 290 km and a five-day revisit time at the Equator (Anon 2019*cde*). Each Sentinel-2 image contains 13 spectral bands encoded to 16-bit per pixel, with spatial resolutions of 10 m (4 bands), 20 m (6 bands), and 60 m (3 bands). Only the 10- and 20-m spatial resolution bands were used for the study. To capture the changes due to seasonal phenology, at least three single, nearly cloud-free Sentinel-2 acquisitions are needed (preferably more), of which one during the leaf-off period (mid-spring) to help distinguish between coniferous and broadleaves, and two during the growth season to help discriminating the tree species in mixed forest stands. Therefore we decided to search for useful imagery data for the time intervals between May-November 2018 and 2019. Sentinel-2 data (geometrically corrected, georeferenced and calibrated to bottom-of-atmosphere reflectance, Anon 2019*de*) were downloaded freely from Copernicus Open Access Hub as 2A-level products using the open-source package 'sen2r' (Ranghetti et al. 2020) of the R statistical software (R Core Team 2020), using a limiting criterion for cloud coverage of maximum 25%. The image bands at 20-m spatial resolution were resampled to 10 m to obtain stacks of congruent rasters, and clipped using the microstand polygon borders. A quality check was performed on each microsegment to remove the image pixels that were not related to vegetation using the

Scene Classification Layer (SCL) which is also provided as a Sentinel-2 Level-2A output (Gascon et al. 2017, Anon 2020*de*). All the microstands included in the analyses have at least 25% of the pixels assigned to vegetation class.

A total of 46 pairs of Sentinel-2 images (an image for each of the 33VWK and 33VXK tiles) were obtained – 28 pairs of images for 2018 and 18 for 2019 (Table 4). In addition, several Sentinel-2 RS vegetation indices (Henrich et al. 2009; Henrich et al. 2012) vegetation such as Chlorophyll Index Green (CIgreen), Chlorophyll Index Red Edge (CIrededge), Enhanced Vegetation Index (EVI), Normalised Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), Transformed Soil Adjusted Vegetation Index (TSAVI), Wide Dynamic Range Vegetation Index (WDRVI), Visible Atmospherically Resistant Indices 700 (VARI700) were also calculated for each microstand from the BOA reflectance (Ranghetti et al. 2020). The formulae describing these indices can be found online at https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/indexdb/.

Table 4. The number of Sentinel-2 (A and B) images by year and by month available for the study area satisfying the condition of maximum 25% cloud coverage.

Year	Month							Total
	May	June	July	August	September	October	November	
2018	7	3	7	2	4	5	0	28
2019	1	4	4	3	4	1	1	18
Total	**8**	**7**	**11**	**5**	**8**	**6**	**1**	**46**

2.5 FOREST STATE MAPS

The raster maps for the main forest state attributes were produced between 2009-2014 using the national ALS survey data. The Swedish Forest Agency (*Skogsstyrelsen*) and the Swedish University of Agricultural Sciences (Nilsson et al. 2017) combined the laser data and field plots from the Swedish National Forest Inventory to produced digital maps of the main attributes characterising the forest state, including basal area (BA), mean basal area weighted diameter and height (Dg and Hg, respectively), and standing volume (VOL), at a spatial resolution of 12.5 x 12.5 m. The map service, "Skogliga grunddata" (SKGD), 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 2019f). Currently, the forest state attributes (VOL, BA, Dg and Hg) provided by SKGD are also some of the most important predictors for yield predictions in the imputation system developed by Skogforsk (Söderberg 2015, Söderberg et al. 2017, 2018). An example of a forest stand with microstand partitioning and selected auxiliary information is presented in Figure 3.

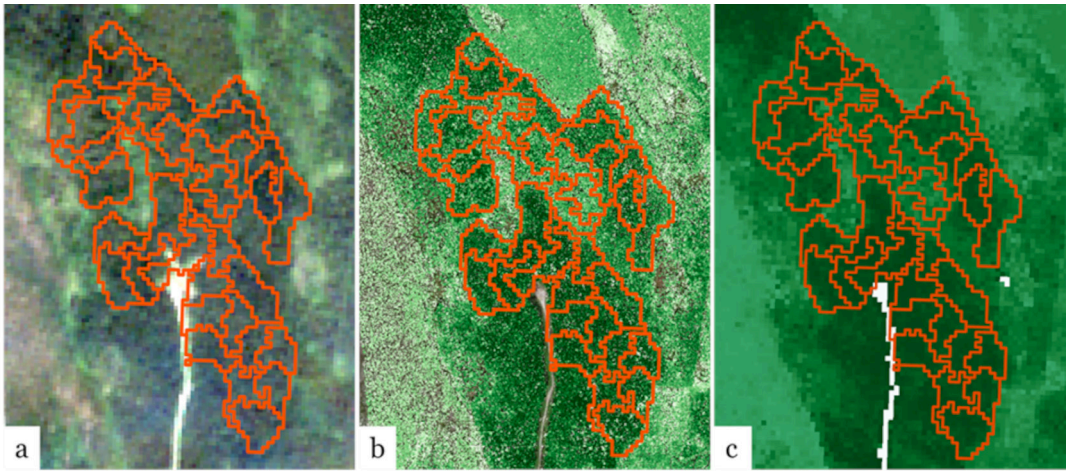

Figure 3. Forest stands and microstands as overlay on true colour (RGB) Sentinel-2 imagery (a), Canopy Height Model (provided by the Swedish Forest Agency) derived from the national airborne laser scanning survey data (b), and volume estimates from Skogliga grunddata (c).

2.6 FEATURE EXTRACTION FROM AUXILIARY DATA

In order to build predictive machine learning algorithms, the microstand-level auxiliary information in raster format has to be represented more compactly to allow ingestion into the machine learning algorithms. Hence, several features were extracted from each type of auxiliary datasets to obtain vector-type data.

The averages and inter-quartile ranges were compiled from the unfolded Sentinel-2 bands and vegetation indices rasters, as well as for each forest attribute from the SKGD rasters. The eigen values of the unfolded raster stacks (i.e., the 10 bands of a Sentinel 2 image) were used to distribute image variability across the 10 radiometric bands. Prior to feature extraction, the imagery data was cleaned up by removing pixels unlikely to cover forest vegetation, using a mask corresponding to CHM regions with an average height below 2 m.

Several descriptors of the vertical distribution of the forest canopy were extracted from the pixels in the CHM rasters above a 2-m height threshold, following Næsset (2004);

- the percentiles corresponding to 0.25, 0.5, 0.75 and 0.95 quantiles of the CHM height distribution within microstands.
- canopy densities derived by first dividing the range between the lowest (>2 m) and the 95th percentile CHM height into 10 equal height intervals. Canopy densities were then computed as the proportion of CHM pixels above each fraction to total number of CHM pixels.

In addition, textural measures (mean, variance, homogeneity, contrast, dissimilarity, and entropy) were extracted from grey-level co-occurrence matrices (Haralik 1973, Figure 2) to characterise the spatial patterns in the CHMs at microstand level. CHM values were quantized to 16 grey levels, and the grey-level co-occurrence matrices (GLCM) were computed using a moving window of 3x3 pixels (to approximate the Sentinel-2 pixel size of 10x10m) over four directions with 90-degree shifts. The computations were performed using the 'glcm'-package (Zvoleff 2020) of the R statistical software (R Core Team 2020).

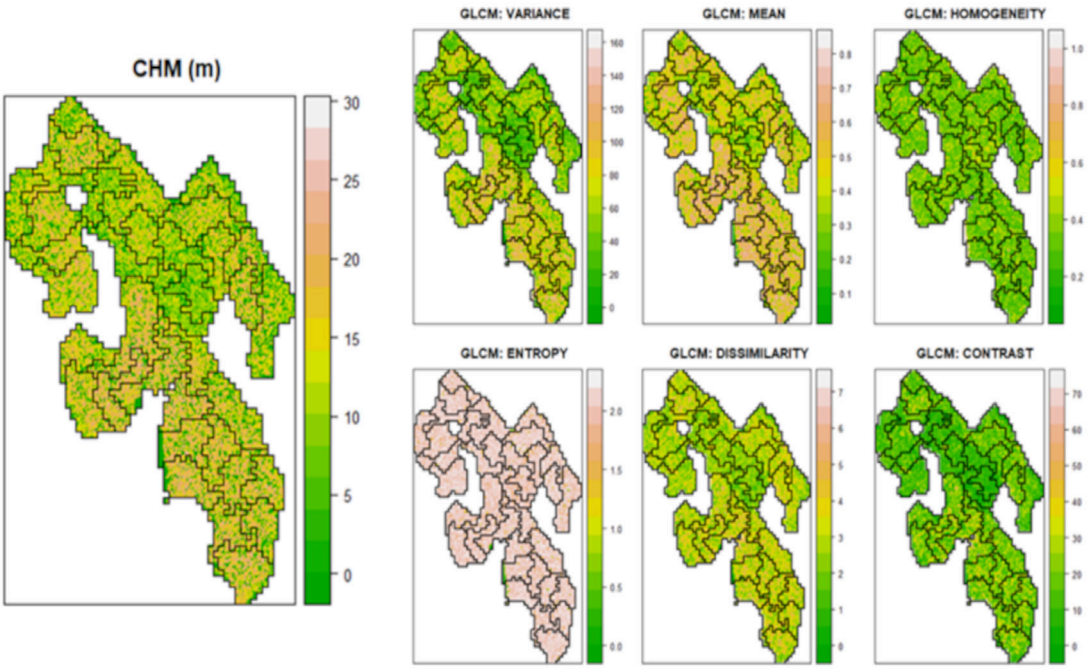

Figure 4. Features extracted from the grey-level co-occurrence matrix (GLCM) describing the height variation patterns in the Canopy Heigh Model (CHM).

3. Statistical methods

The main steps in method development are described in the following sections. The procedures were run using a data splitting approach, by randomly splitting the ground-truth dataset into training-testing (75%) and validation (25%) datasets. The sampling units for data splitting were the forest stands, to preserve the hierarchical structure of the data characterised by strong dependencies among the microstands within the same forest stand. Splitting the data by microstand would lead to overoptimistic results, since microstands from the same objects may end up in both the training-testing and the validation datasets.

3.1 PREDICTIVE MODELS FOR TREE SPECIES PROPORTIONS

Tree species mapping requires that ground-truth observations (compiled from harvester data, see section 2.1) be linked to the vector of features extracted from auxiliary data (section 2.5). The result is a multi-temporal dataset containing the ground-truth observations (which do not vary over time) and the set of auxiliary information obtained at each of the 46 time points between May-Nov. 2018 and 2019. In this dataset, the entire vector of imagery-related auxiliaries can be missing for several microstands at different time points, due to quality requirements imposed for imagery data (section 2.3). The missing auxiliary data will also produce a large variation in the sample sizes (approx. 43% for training-test and 48% for the validation data, Table 5) containing both ground-truth observations (microstands) and imagery data along the two-year interval.

Table 5. Sample size variation for training and validation temporal datasets.

Datasets	Statistic		
	Range (min-max)	Average	CV (1)
Train	288-1781	1341	43.3
Valaidation	45-299	254	47.6

(1) Coefficient of variation

There are several modern approaches addressing cloud removal and land surface reconstruction using satellite data (see Shen et al. 2015, Meraner et al. 2020 and references therein), but they require large datasets, and the final result is an altered image product. Here we adopted a different strategy for handling missing imagery data in multi-temporal datasets, one that requires less data and does not alter the image content.

Tree species mapping was addressed as the process of predicting species proportions, and not as a traditional classification approach based on class probabilities (or odds). The reasoning for this was that predicted tree species proportions are required for distributing the predicted total volumes by species. Accurate species-specific volume predictions can then be used as new auxiliary data for controlling the imputations. Moreover, the assignment into various classes (i.e., the classification) can be seen as a biproduct of predicting the species proportions. For instance, classification into dominant tree species (by proportion in volume) can be easily retrieved after predicting the volume proportion for each species.

The tree species proportions were defined as the portions of the total volume (or total amount) allocated to each individual species (or component) at microstand level. The result is a typical compositional dataset (Aitchison 1982), where each observation (a data point) represents a vector on a simplex (Pawlowsky-Glahn & Egozcue 2001, Billheimer et al. 2001, Pawlowsky-Glahn 2003), as shown in Figure 5:

Figure 5. Ternary diagram for composition data. The data points in the simplex represents the tree species proportion in volume for spruce, pine and deciduous, by microstands.

The analysis of compositional data requires specific statistical methods that can account for the effects introduced by the compositional constraint.

A flexible approach is the use of so-called α-transformation (Tsagris et al. 2011) that extends the log-ratio transformations and naturally handles zero-valued components. The transformed data is mapped into an unconstrained d-dimensional space (d=D-1, where D is the number of components in the original composition) via multiplication with a Helmert sub-matrix. Feasible values for the α parameter ($\alpha \in (-1,1)$) can be estimated from the data and, as for the log-ratio transform, the results can be then mapped back to the simplex via an inverse α-transformation. The α-transformation approach was pursued further, and the compositional data analyses were performed using the 'Compositional'-package (Tsagris & Athineou 2020) of the R statistical software (R Core team 2020).

The α-transformation was first applied to the species proportions (D=3) in the training data, resulting in a bivariate dataset (d=2) of new observations, and the transformed dataset was then analysed further using regression modelling tools. Using the R-package 'MASS' (Venables W.N & Ripley B.D 2002), two sets of predictors were selected among the auxiliary training data using forward stepwise feature selection based on Bayesian

Information Criteria. Predictive models for the transformed variables were then multivariate additive regression models using the R-package 'mgcv' (Wood 2004, 2011). Due to the hierarchical structure of the data (microstands within stands), a stand-level random effect was included in each model as a penalised regression component to account for within-stand variability. This step is important if uncertainty estimates in the form of prediction errors are required. If only point predictions are sought, including random effects in the models, the step may not be necessary. The multivariate predictions were then back-transformed to the simplex.

3.2 SELECTING RELIABLE MULTI-TEMPORAL PREDICTIONS FOR SPECIES PROPORTIONS

The models used for tree species mapping (section 3.1.1) were fitted independently on each time point in the training dataset, using only the observations comprising both ground-truth data and valid auxiliaries available at a particular date. The downside of this approach is that having the training data subsets varying by size hinders a direct comparison of the models. For this reason, the predictive power of the models was assessed using a Leave-One-Out Cross-Validation (LOOCV) resampling procedure following the recommendation of Hastie et al. (2009, §7.10). LOOCV was performed at stand-level by removing all microstands associated with a forest stand at a time to comply with the hierarchical structure of the datasets with microstands nested within stands.

The mean absolute deviations (*mad.cv*) at microstand level and the multivariate predictions resulted after LOOCV were then back transformed to the simplex. The 95th percentile of the conditional *mad.cv* distribution was modelled using additive quantile regression via the R-package 'qgam' (Fasiolo et al. 2017) with the Chlorophyll Index Red Edge (*CIrededge*) vegetation index as covariate. Quantile regression enables the prediction of the entire conditional distribution of the response (Koenker & Bassett 1978). For a specific quantile q, the quantile regression line will split the population scatter in a way that approximately $100q$% of the data points will be located below the line and $100\ (1-q)$% above the line. To guarantee predictions that are bounded to (0, 1), the logistic transformation was applied to *mad.cv* values prior to fitting the quantile regression models, and the predictions (*mad.cv.qu*) were back-transformed. Since the logistic transformation is monotonic and there are no distributional assumptions on which the quantile regression is based, there is no reason to suspect a back-transformation bias. In the case of *mad.cv*, which is bounded to (0, 1), the conditional quantile regression predictions can be seen as approximations for the upper limit of simultaneous ('across-the-function') prediction intervals. This procedure does not guarantee exact coverage for a new dataset, being only a model-based approximation that may suffice for our analyses.

Both criteria (*mad.cv* and *mad.cv.qu*) can be used for assessing the predictive power of the models fitted to multi-temporal data, but *mad.cv.qu* has the advantage of describing uncertainty for the validation data directly.

In addition, dominant tree species (by volume) were classified using LOO-CV at both microstand and stand level. Cohen's Kappa (Cohen 1960) and the macro-averaged F1-score (Optitz & Burst 2019) derived from the confusion matrices using the R-package 'caret' (Kuhn 2020) were used as validation metrics for classification. The goal is to identify the classifiers that maximise each of these criteria. For instance, Cohen's Kappa values <0 can be interpreted as no agreement between the predicted and true classes,

0.01-0.20 as poor classifier performance, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1.0 almost perfect agreement (Landis & Koch 1977). The F1 score varies from 0-1, and it should be as high as possible when averaged over all the classes.

A summary of the scoring criteria characterising the multi-temporal species proportion predictions and the subsequent classification by dominant species is presented in Table 6.

Table 6. Scoring criteria for species proportion predictions and classification by the dominant class obtained on the training-validation datasets and LOO-CV.

Statistic	Criterion									
	F1		Kappa		mad.cv		mad.qu		PI coverage (2)	
	Stand	Micro stand	Stand	Micro stand	Stand	Micro stand	Stand	Micro stand	Stand	Micro stand
Training dataset										
min	0.68	0.57	0.37	0.20	0.08	0.07	1.50	2.20	0.95	0.95
max	0.90	0.90	0.79	0.79	0.13	0.13	2.51	3.46	0.99	0.96
mean	0.79	0.81	0.57	0.62	0.11	0.10	1.97	2.83	0.96	0.95
CV (1)	7.31	8.28	19.71	20.50	10.64	12.65	14.45	10.11	0.90	0.20
Training+LOOCV										
min	0.67	0.50	0.35	0.07	0.08	0.08	2.15	2.52	0.92	0.86
max	0.90	0.89	0.79	0.78	0.13	0.13	3.83	4.39	0.97	1.00
mean	0.78	0.79	0.56	0.58	0.11	0.11	2.77	3.39	0.95	0.93
CV (1)	7.25	9.81	19.83	24.70	12.34	12.34	13.97	13.73	1.07	2.91
Validation dataset										
min	0.65	0.52	0.21	0.08	0.08	0.08	1.51	2.19	0.75	0.84
max	0.95	0.94	0.84	0.86	0.16	0.14	2.76	3.44	0.95	0.98
mean	0.85	0.84	0.62	0.64	0.12	0.11	1.99	2.86	0.88	0.93
CV (1)	7.67	10.78	56.83	50.82	15.29	15.78	14.26	9.95	5.97	3.50

(1) Coefficient of variation in percentages relative to average

(2) Actual coverage of the simultaneous prediction intervals; the nominal coverage is 95%

The average values of *mad.cv* and *mad.cv.qu* at microstand level resulted from LOO-CV were then aggregated at each time point in the multi-temporal dataset. Each microstand in the validation data was then assigned the predictions produced by the best ranking model according to these two criteria.

3.3 UPDATING THE FOREST STATE ESTIMATES

The yield prediction method implemented at Skogforsk (Söderberg et al. 2017, Söderberg et al. 2018) is based on imputations using SKGD data (VOL, BA, Dg and Hg) as auxiliary information. However, SKGD can be outdated in certain situations due to the time interval or to major forest disturbances that occurred after the ALS data acquisition. Although we are rather confident that the latter situation is not a major concern for our analyses, we decided to update the forest state estimates in the SKGD by deploying statistical models developed from our sample data. Intuitively, it is expected that using more accurate forest state estimates would increase the imputation accuracy as well. Generalised additive regression models assuming a Gaussian distribution of the response and a square root link-function were fitted on the training data following the same steps as in section 3.1.1. The scatterplots of predicted versus ground-truth forest attributes in the validation data presented in Figure 6 indicate that the predictions based on updated (panel b) SKGD data are more accurate.

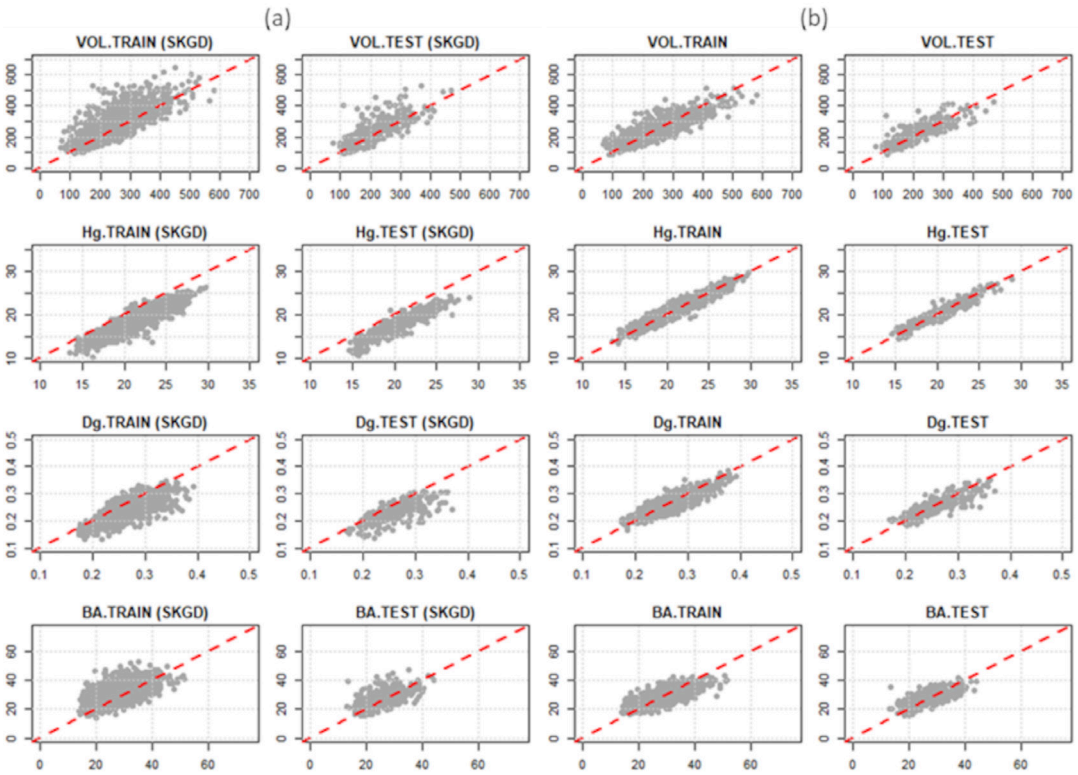

Figure 6. Predicted versus ground-truth for forest state attributes on training and test datasets. Panel a: ground-truth forest state estimates (x-axis) versus Skogliga grunddata (SKGD). Panel b: ground-truth forest state estimates (x-axis) versus forest state attributes predicted using models fitted to sample data.

The species proportions predicted using the methods presented in section 3.1.1 were multiplied with the total volume predictions to produce the species-specific volumes, as shown in Figure 7.

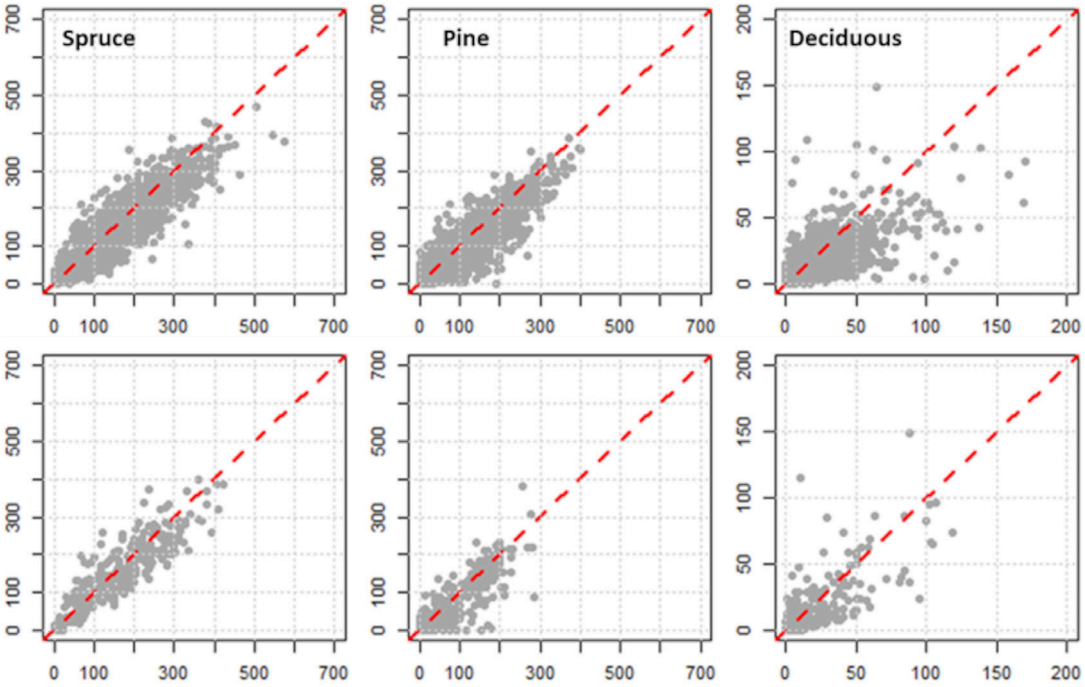

Figure 7. Species-specific volumes (m^3/ha) obtained from regression estimation. The results for training data are on the upper row, and the bottom row contains the results for the validation dataset.

3.4 YIELD PREDICTIONS VIA NEAREST-NEIGHBOUR IMPUTATION

Nearest-neighbour imputations are simple, greedy algorithms used for searching in the feature space for finding the k (k≥1) observations that are close to each other based on some type of similarity measure. For instance, having a data sample (reference data) of multivariate observations containing ground-truth and auxiliary information, the imputation method will assign (or impute) the ground-truth data to new observations (target data) for which only the auxiliary data is available, such as forest stands that are scheduled for harvesting. Besides their relative simplicity, a major advantage of the imputation methods is the ability to handle multivariate ground-truth observations. Imputing several forest attributes from the reference data simultaneously can help preserve their natural covariance structure.

Imputing several reference observations (i.e., k>1) to a target observation can help increase the accuracy, but the accuracy deteriorates for large k values. The imputations for finding feasible k-values are usually tuned using a cross-validation approach. Based on results from previous projects, we adopted k=5 for our analyses following the approach presented in Söderberg et al. (2018) and Söderberg et al. (2019) that is using the Most Similar Neighbour method (Moeur & Stage, 1995). The computations were performed using the open source 'yaImpute' package (Crookston & Finley 2007) of the R statistical software (R Core Team 2020).

The auxiliary data for imputations consisted of (1) forest state estimates (VOL, BA, Dg and Hg from SKGD), and (2) forest state estimates (BA, Dg and Hg) predicted by the local models developed on the data training data subsets in combination with species-specific volume predictions, as described in section 3.1.1.

4. Results and discussion

The yield prediction assessments were run under three scenarios:

A. Baseline, with no species information incorporated in the imputations. Although some information on tree species distribution may exist, it is not always available and is not spatially explicit. In addition, at microstand level, such data is relevant only if the species composition is strongly dominated by one tree species. Consequently, the baseline can be seen as the worst-case scenario.

B. Using tree species predictions as predicted species proportions (B.1), hot encoded as indicator data (0-1) by true dominant species (B.2) and predicted dominant species (B.3) by volume.

C. Using 'perfect' information (from the 'ground-truth' data) on species-specific volumes.

For each scenario, the average individual tree volumes by DBH class (Table 1) were calculated using the tree lists obtained from the harvester production files, and the average tree species-specific yield values were obtained as SEK/ha. The species-specific gaps in monetary value are presented in SEK/ha and %, and the relative results (%) including the overall gaps are summarised in Table 7 and Table 8:

Table 7. Gaps (%) to the ground-truth values under the three scenarios produced by imputations, using Skogliga grunddata forest state estimates and predicted species-specific volumes.

Species	Statistic	Scenario				
		A	B.1	B.2	B.3	C
Spruce	Mean	49.05	24.10	31.19	35.11	9.03
	CV (1)	15.83	8.14	9.27	12.15	19.42
Pine	Mean	89.39	41.88	49.70	61.86	13.10
	CV	16.53	19.61	17.56	21.93	19.12
Deciduous	Mean	75.67	59.53	69.95	78.61	14.20
	CV	7.53	9.30	7.94	24.77	15.36
Overall	Mean	18.16	14.14	17.84	17.82	4.14
	CV	6.31	5.18	6.18	6.70	8.59

(1) Coefficient of variation (%)

Table 8. Gaps (%) in the ground-truth values under the three scenarios produced by imputations, using updated forest state estimates (BA, Dg and Hg) and predicted species-specific volumes.

Species	Statistic	Scenario				
		A	**B.1**	**B.2**	**B.3**	**C**
Spruce	Mean	37.85	25.42	25.82	29.76	9.90
	CV [(1)]	14.54	8.95	11.83	13.25	15.91
Pine	Mean	71.62	43.19	44.87	58.06	14.87
	CV	19.13	19.85	17.90	24.54	20.82
Deciduous	Mean	76.58	59.04	71.37	80.42	16.39
	CV	8.49	11.05	7.14	23.87	11.72
Overall	Mean	14.41	14.06	13.99	14.19	4.63
	CV	5.21	4.52	5.00	5.50	9.17

[(1)] Coefficient of variation (%)

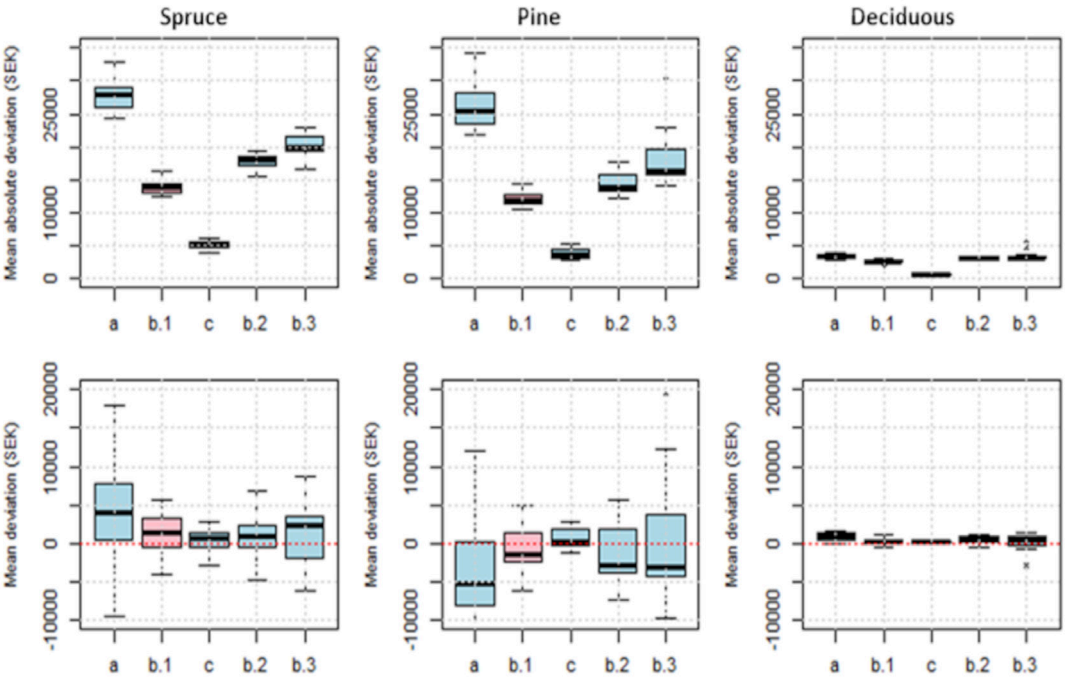

Figure 8. Species-specific gaps (SEK) under the three scenarios considered for imputations using forest state estimates from Skogliga grunddata in combination with regression-based species-specific volumes: A - no species information; B - predicted species information in the form of species proportion (B.1), predicted dominant species classes (B.2), true dominant species classes (B.3), and the 'perfect information' case (C). The mean absolute differences are shown in the upper row, and the mean differences are in the lower row.

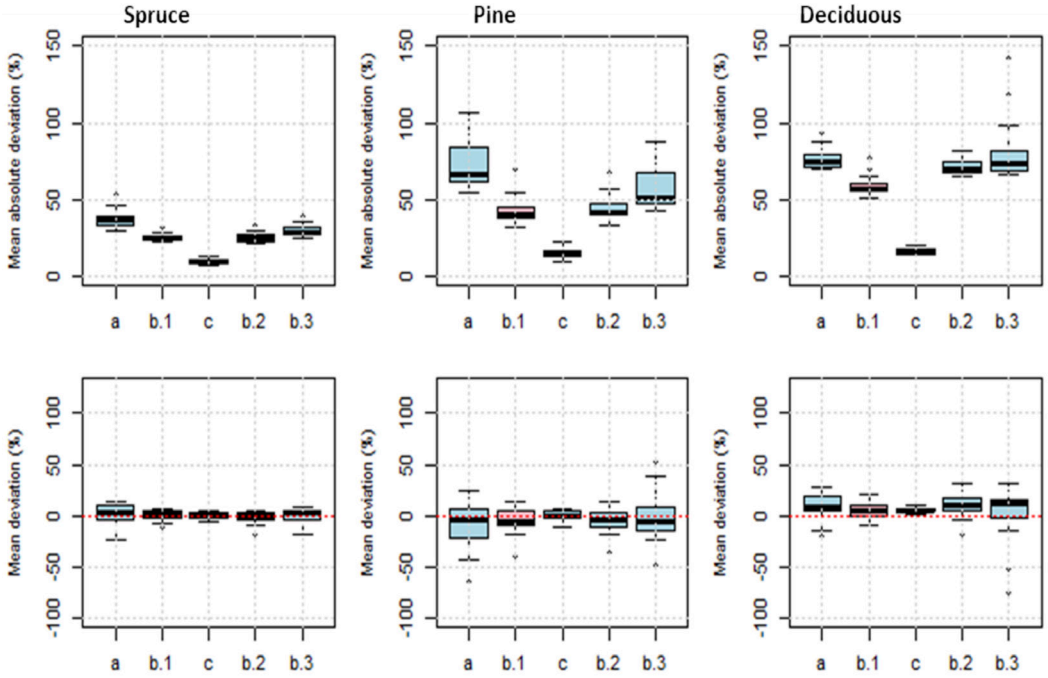

Figure 9. Species-specific gaps under the three scenarios considered for imputations using updated forest attributes as auxiliary data in combination with regression-based species-specific volumes: A - no species information; B - predicted species information in the form of species proportion (B.1), predicted dominant species classes (B.2), true dominant species classes (B.3), and the 'perfect information' case (C). The mean absolute differences are shown in the upper row, and the mean differences are in the lower row.

The classification results (Table 6) are heavily influenced by unbalanced training data, where the Deciduous species constituted a very small proportion. However, Cohen's Kappa indicates a good to excellent agreement beyond chance between true and predicted classes (Fleiss et al. 2003, §18), and the average F1-scores around 0.8 are rather high (the closer to 1, the better). We would expect better results for a training dataset containing a larger sample of deciduous species. The average actual coverage of the simultaneous predictions intervals was slightly too short, missing the nominal 95% coverage by approximately 2 percentage points.

The methodology cannot cope with microstands for which there is no imagery data available during the specific time interval. Possible solutions for such cases may be searching for imagery data within a wider time interval, using different source of auxiliary data such as radar imagery, relying exclusively on the auxiliary information provided by textural descriptors of the CHM models, or performing yield predictions without incorporating species information. For the latter case, the species composition at forest stand level can be retrieved from pre-harvesting field inventories. Alternatively, the stand-level information on species distribution from forest registers can also be used, but a loss in accuracy should be expected, since information quality is not always reliable and does not reflect local variability at microstand level.

The project demonstrates that the value of species information can be successfully quantified in economic terms. Even in regions with a small number of tree species, the predicted species proportions can noticeably improve the value recovery predictions. The best results were achieved when predicted tree species information in the form of species-specific volumes (m^3/ha) was used in the imputations. The loss reduction was about 100% for Spruce and Pine, and approximately 20% for Deciduous species. This is in line with the species prediction accuracy, which was lowest for the minority class (Deciduous species).

The gaps to the ideal scenario C were also smallest when using predicted tree species proportions (scenario B.1). When combined with the default SKGD information (Table 7), the gap for Spruce and Pine decreased by 59-62 percentage points (pp) compared to the baseline (scenario A), and 19-42 pp relative to scenarios B2 and B3, while for Deciduous the gap was reduced by approximately 24 pp (scenario A) and 17-27 pp under scenarios B2 and B3. Updating the SKGD auxiliaries (Table 8) improved the accuracy under all scenarios, but the gaps remained substantial, up to 44-50 pp for Spruce and Pine and about 34 pp for Deciduous. The overall loss reduction was approximately 22%, which is much less compared to the species-specific results. This is explained by the random compensations between over-and-underestimated yield predictions that can cancel each other out to a certain extent. However, the lower gap reduction due to random cancellations does not appropriately characterise the results, since supply chain optimisation and industry require information on species-specific wood assortments.

The results also suggest that incorporating tree species information as predicted volumes by species is more efficient than using binary hot encoding for the dominant species. Although the classification is a by-product of the species proportion predictions, using continuous auxiliaries (volume estimates) instead of categorical (class labels) seems to improve the search results during imputations.

The gains in the predicted yield are directly dependent on two factors:

(1) the accuracy of total volume estimates, and

(2) the accuracy of species proportion predictions.

The results also indicate that the information on species proportions does not necessarily help improve the accuracy for total quantities (i.e., volumes m^3/ha). However, more accurate volume estimates increase the accuracy of species-specific volume estimates.

5. Conclusions

Although the project addresses only a very specific link in the wood supply chain, it is expected that the gains generated by increased yield prediction accuracy will be compounded along the supply chain. More accurate yield predictions will also help increase efficiency in various planning activities related to transport and logistics and secure a better raw material supply for industry.

The methodology can be further improved by extending the research in several directions, such as;

- by including study areas with a higher tree species diversity.
- by ensuring a more balanced dataset with regard to species distribution, to avoid including minority classes in the analyses.
- by including vegetation height raster maps from multi-temporal airborne laser scanning data. A new data acquisition campaign started in 2018, and will cover 75% of the Swedish area (approx. 350,000 km2) with 1-2 points m^2 (Anon 2019d). The differences in average forest height between two time points would reflect the site productivity. If the forest age can be retrieved from the local data (like forest registers), the estimated change can be linked to species-specific site productivity for a finer tuning of the imputation method.
- by developing a procedure for using the vegetation classes from the upcoming National Land Cover Database (currently under development) maintained by the Swedish Environmental Protection Agency as proxies for estimating tree species proportions and for mitigating missing data issues related to cloud occlusions.
- not least, including information on wood quality compiled from the harvester production files would enable a more comprehensive assessment.

References

Anon 2020a. Lantmäteriet. Product description: Laser data - Laserdata Skog. Available online at: https://www.lantmateriet.se/sv/Kartor-och-geografisk-information/Hojddata/GSD-Hojddata-grid-2/, last accessed on 14 January 2020.

Anon 2020b. Skogsstyrelsen: Trädhöjdsraster laser – produktbeskrivning . Available online at: (https://www.skogsstyrelsen.se/globalassets/sjalvservice/karttjanster/geodatatjanster/produktbeskrivningar/raster-tradhojd-laserdata-nh---produktbeskrivning.pdf, accessed on 14 January 2020)

Anon 2019c. ESA. Sentinel satellites - Overview. Observing the Earth. Available online at: http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview4, last accessed on 07 January 2020.

Anon 2019d. European Space Agency. Sentinel-2 User Handbook. Available online: https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook, last accessed on 09 January 2020.

Anon 2020e. European Environment Agency: Level-2A Algorithm Overview. (URL: https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm, last accessed on 14 January 2020)

Anon 2019f. Skogsstyrelsen (URL: https://www.skogsstyrelsen.se/skogligagrunddata, last accessed on 14 January 2020)

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20, 37-46 (https://doi./10.1177/001316446002000104)

Crookston. N.L. & Finley, A.O. 2007. yaImpute: An R Package for k-NN Imputation. Journal of Statistical Software 23, 1-16.

Fasiolo, M., Goude, Y., Nedellec, R. & Wood, S.N. 2017. Fast calibrated additive quantile regression. URL: https://arxiv.org/abs/1707.03307

Fleiss, JL., Levin, B. & Paik, M.C. 2003. Statistical Methods for Rates and Proportions (3rd ed.). Hoboken, NJ: Wiley.

Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J. et al. 2017. Copernicus Sentinel-2A Calibration and Products Validation Status. Remote Sensing, 9, 584 (https://doi.org/10.3390/rs9060584)

Haralick, R.M., Shanmugam, K. & Dinstein, I. 1973. Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3, 610-621

Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer.

Henrich, V., Krauss, G., Götze, C., Sandow, C. 2012: IDB - www.indexdatabase.de, Entwicklung einer Datenbank für Fernerkundungsindizes. AK Fernerkundung, Bochum, 4.-5. 10. 2012.

Henrich, V., Jung, A., Götze, C., Sandow, C., Thürkow, D. & Gläßer, C. 2009: Development of an online indices database: Motivation, concept, and implementation. 6th EARSeL Imaging Spectroscopy SIG Workshop Innovative Tool for Scientific and Commercial Environment Applications Tel Aviv, Israel, March 16-18, 2009.

Koenker, R. & Bassett, G.W. 1978. Regression Quantiles. Econometrica, 46, 33-50.

Kuhn, M. 2020. caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

Landis, J.R. & Koch, G.G. 1977. The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Meraner, A., Ebel, P., Zhu, X.X. & Schmitt, M. 2020. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS Journal of Photogrammetry and Remote Sensing, 166, 333-346 (URL https://doi.org/10.1016/j.isprsjprs.2020.05.013)

Moeur, M. & Stage, A.R. 1995. Most similar neighbor: An improved sampling inference procedure for natural resource planning. Forest Science, 41, 337-359.

Næsset, E. 2004. Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scandinavian Journal of Forest Research, 19, 164-179

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/).

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.

Ranghetti, L., Boschetti, M., Nutini, F. & Busetto, L. 2020. sen2r: An R toolbox for automatically downloading and preprocessing Sentinel-2 satellite data. Computers & Geosciences, 139, 104473. DOI: 10.1016/j.cageo.2020.104473, URL: http://sen2r.ranghetti.info/.

Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li,H. & Zhang, L. 2015. Missing information reconstruction of remote sensing data: a technical review IEEE Geoscience Remote Sensing. Magazine, 3, 61-85

Söderberg, J., et al. 2017. Utvärdering av utbytesprognoser med skogliga laserskattningar och skördardata - Evaluation of yield forecasts produced by forest laser estimations and harvester data. Work Report 937-2017. Uppsala, Skogforsk.

Söderberg, J., Möller, J. & Willén, E. 2018. Evaluation of yield prediction with harvester data. Work Report 981-2018, Skogforsk (in Swedish).

Tsagris M.T., Preston, S. & Wood, ATA. 2011. A data-based power trans-formation for compositional data. In Proceedings of the 4th Compositional Data Analysis Workshop, Girona, Spain.

Tsagris M & Athineou G (2020). Compositional: Compositional Data Analysis. R package version 4.2. https://CRAN.R-project.org/package=Compositional

Venables, W.N. & Ripley B.D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wood, S.N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association, 99, 673-686.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B), 73, 3-36.

Zvoleff, A. 2020. glcm: Calculate Textures from Grey-Level Co-Occurrence Matrices (GLCMs). R package version 1.6.5. https://CRAN.R-project.org/package=glcm