DE GRUYTER
OPEN

Estimation of standing wood volume and species composition in managed nemoral multi-layer mixed forests by using nearest neighbour classifier, multispectral satellite images and airborne lidar data

Mait Lang[1,2*], Tauri Arumäe[2], Tõnu Lükk[1,2] and Allan Sims[2]

Lang, M., Arumäe, T., Lükk, T., Sims, A. 2014. Estimation of standing wood volume and species composition in managed nemoral multi-layer mixed forests by using nearest neighbour classifier, multispectral satellite images and airborne lidar data. – Forestry Studies | Metsanduslikud Uurimused 61, 47–68. ISSN 1406-9954. Journal homepage: http://mi.emu.ee/forestry.studies

Abstract. Nearest neighbour techniques are useful for constructing maps of forest inventory variables based on sample plot and auxiliary data from remote sensing. The most problematic issue of the nearest neighbour technique is possible systematic bias in the estimated values. In this study a 15 by 15 km test site in nemoral multi-layer mixed forests was established in Laeva, Estonia. A set of 444 circular sample plots was used as reference set. Airborne lidar data and Landsat-8 Operational Land Imager image were used to construct five different feature variable sets consisting of original variables and alternatively principal components. The response variables were wood volume of first tree layer, wood volume of the second tree layer and main species code. A special test was carried out where a substantial amount of Silver birch dominated plots were removed from the reference set. The wood volume prediction validation was carried out on 89 forest growth sample plots and on 2290 forest stands. Species composition prediction was validated on 986 forest stands. As in many previous studies the results confirmed superiority of airborne lidar variables over spectral variables for wood volume estimation. The first three principal components of airborne lidar variables and first five principal components of all possible original feature variables contained over 99% of the information and performed well in imputations. The imputed wood volume at small values was overestimated and underestimated at large values regardless of used reference set. The feature variable sets containing spectral data performed better for species composition imputation. There was a forest age dependent discrepancy in predicted species proportions: birch and spruce proportions were underestimated in young stands and overestimated in older stands while proportion of aspen had exactly the opposite errors. The lack of fit depended slightly on the feature variable sets. The birch dominated plot partial removal from the reference set changed the predicted proportion of species but did not remove the forest age dependent lack of fit. The result can be important for the studies in which bootstrap samples are used to estimate error statistics for nearest neighbour technique based forest inventory variable maps.

Keywords: nearest neighbour technique, spectral data, lidar data, mixed forest, wood volume, species composition.

Authors' addresses: [1]Tartu Observatory, 61602 Tõravere, Tartumaa, Estonia; [2]Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Kreutzwaldi 5, Tartu 51014, Estonia; *e-mail: lang@emu.ee

DOI: 10.2478/fsmu-2014-0010

Introduction

Sustainable forest management and forest policy implementation can only be based on timely updated and accurate estimates of forest resources. The elementary observation units can be forest stands which are homogeneous parts of forest delineated for common silvicultural treatments or instrumentally measured sample plots which are further used to produce statistical estimates for a stand, a small area, a region or a country (Krigul, 1972). Such a regularly distributed plots based sampling is used for National Forest Inventories in many countries including Estonia (Tomppo *et al*., 2010; Adermann, 2010). Several data linking methods and sampling designs have been tested to incorporate remote sensing data i.e. aerial photos, multi spectral satellite images, radar data and airborne laser scanning data to improve the accuracy of regional estimates and to construct updated maps of forest inventory variables (Poso *et al*., 1990; Howard, 1991; McRoberts & Tomppo, 2007). One of the most popular techniques has been nearest neighbour imputation (Fazakas *et al*., 1999; Holmgren *et al*., 2000; Huiyan *et al*., 2006; McRoberts & Tomppo, 2007; Packalén & Maltamo, 2007; Chirici *et al*., 2008; Kajisa *et al*., 2008; McInerney & Nieuwenhuis, 2009; Breidenbach *et al*., 2010; McRoberts, 2012; Fassnacht *et al*., 2014; Zald *et al*., 2014). In Estonia, Tamm & Remm (2009) used stand-wise forest inventory data and nearest neighbour technique based machine learning algorithms to construct forest maps of inventory variables.

The *k* nearest neighbour imputation (*k*NN) for forest inventory is based on the assumption that the response variables i.e. the variables measured on elementary observation units in a reference set can be assigned to a target set by using a similarity measure based on auxiliary predictor variables i.e. feature variables which are obtained mainly from remote sensing. The elementary sampling units of feature variables are usually raster image pixels. The reference set elements are instrumentally measured and georeferenced sample plots with location coordinates linking the plot data to the observations of feature variables. In nearest neighbour imputation the target set pixels are assigned weighted averages of response variables of *k* nearest observations from reference set where for each observation both the response variables and feature variable values are known (Fazakas *et al*., 1999; McRoberts, 2012). The Euclidean distance is the most common metric for estimating similarity of a target set pixel to a reference set observation.

According to McRoberts (2012) the nearest neighbour techniques are appealing since they can be used for map construction, they can handle continuous and categorical response variables, there are no assumptions regarding the distributions of response or predictor variables, and they can be applied for a wide range of data sets. On the other hand, the main concern about spectral feature variables based nearest neighbour techniques for forest inventory is the possible bias in estimates (Poso *et al*., 1990; Holmgren *et al*., 2000) which is caused by the nonlinear and saturating relationships of the feature variables to response variables, e.g. age and spectral reflectance of forest (Nilson & Peterson, 1993), and the fact that all weights are positive (Fazakas *et al*., 1999). When the observations of selected nearest neighbours are near their smallest or largest values under- or overestimation can occur. However, if the target set estimates are aggregated for an area of several hundreds of hectares the mean values are assumed to be unbiased and the root mean square error of estimates decreases (Fazakas *et al*., 1999). Holmgren *et al*. (2000) describe a systematic lack of fit in multispectral satellite images based *k*NN (*k* = 5) imputed stem wood volume estimates at stand level (overestimation at small values and underestimation at large values) and propose a linear correction model to decrease the estimation error. Kajisa *et al*. (2008) used *k*NN in conifer plantations in

Japan and also indicated similar systematic lack of fit in the imputed stem volumes at sample plot level when $5 \leq k \leq 10$. However, McRoberts (2008, 2012) shows that forest stand level estimates of wood volume, basal area and number of trees per unit area can be obtained with reasonable accuracy if optimal set of feature variables is selected and k value is optimized during the imputation procedure. On the other hand, Gilichinsky *et al*. (2012) propose histogram matching of *k*NN imputed wood volume to the corresponding statistics from field inventory to decrease the lack of fit and reduce estimation errors. However, Gilichinsky *et al*. (2012) did not study the impact of this procedure on other response variables.

The nearest neighbour methods can handle arbitrary set of feature variables, but the unrelated ones to the response variables increase uncertainty of estimates (McRoberts, 2008). Reduction of feature variable space can be accomplished by using genetic algorithms to select the most informative set of feature variables (McRoberts, 2008; Tomppo *et al*., 2009), cluster analysis (Tamm & Remm, 2009), principal components or spectral variable transformation to vegetation indices (Chirici *et al*., 2008), or an expert guess (Zald *et al*., 2014). Latifi & Koch (2012) used evolutionary genetic algorithm to select 12 covariates from initial 68 predictor variables calculated from full-wave ALS data and four band multi spectral line scanner data. The estimates of biomass, number of stems per unit area and standing volume for target area were imputed by using regression tree method and *k*NN with the canonical correlation analysis distance metric (*k*MSN) which basic principle principle was described by Moeur & Stage (1995). The regression tree imputation method was suggested in case of large number of well correlated predictor variables. However, the *k*MSN imputed stem volume estimates had smaller overall bias. Latifi & Koch (2012) also showed that both imputation methods have systematic lack of fit in stand level estimates of stem volume.

One important issue regarding the nearest neighbour techniques is related to the number of nearest neighbours, i.e. the value of k. Based on earlier studies, Fazakas *et al*. (1999) conclude that increase in k improves wood volume estimation accuracy up to $5 \leq k \leq 10$ spectral neighbours, with the price of artificially improved correlation between response and feature variables compared to the field measurements. It is the characteristic of *k*NN that increase in k shifts the estimated values towards the sample mean and the rate of decrease of prediction errors can be used to select optimal k value. Chirici *et al*. (2008) indicate that increase of k over 5 does not substantially decrease estimation errors. However, McRoberts (2008, 2012) shows that genetic algorithm based optimization can yield values of $25 \leq k \leq 47$ and McRoberts (2012) shows that k is related to the distance-based weight t imposed on the neighbours used in the calculation of predictions.

The aim of the study was to test the performance of a simple *k*NN imputation technique in managed nemoral multi-layer mixed forests to estimate 1) standing wood volume for first (dominant) and second (lower) tree layer, and 2) to estimate main species for target pixels and species composition for forest stands. The feature variables were extracted from Landsat-8 OLI image and from airborne laser scanning data. Five combinations of feature variables were tested. The reference set consisted of 444 sample plots and validation was carried out on exclusive 89 forest growth network plots and on 2290 forest stands. A special test was carried out where 70% of Silver birch dominated reference plots removed from the reference set to study the stability of species composition prediction.

Material and methods

Test site

The 15 by 15 km test site (Figure 1) is located in southern Estonia (centre coordi-

nates in EPSG:3301 projection: 6490854 N; 642472 E) and the area is mostly dominated by mixed multi-layered deciduous forests. Dominating tree species are Trembling aspen, Silver birch, Norway spruce, Grey alder and Black alder. There are also Scots pine stands growing on some less fertile sites. The second i.e. the lower layer in the deciduous forests is usually dominated by shade tolerant Norway spruce. The *Padus avium* Mill. or *Corylus avellana* L. dominated forest understory had canopy cover in many plots over 20% (Appendix 1). There are mainly two different forest site types according to Lõhmus (2004) – *Aegopodium* (AG) and *Filipendula* (FP). The soil types are according to FAO-UNESCO mainly either *Calci Eutric Gleysols* or *Eutri Histic Gleysols*. Most of the forests are managed actively.

Figure 1. Test site location map.

Joonis 1. Testala asukoht.

Sample plots and forest inventory data

First step to create *k*NN reference set was to use stand-wise forest registry data to find the most frequently occurring dominant tree species in Laeva test site area. Then 500 stands were randomly chosen to represent species and age distribution (young, middle aged, old-grown) of the test site. Only the stands with area over 1 ha were sampled. The radius for reference sample plots was 10 m and the plots were placed in a homogeneous and representative area in each selected stand, with at least 20 m inside from the edge. All the trees with diameter at breast height larger than 4 cm were measured with calliper and a minimum of 15 trees was selected for height measurements. In very young stands the stand density, canopy cover, mean tree diameter at breast height and forest height were estimated in the field. Table 1 provides the general description and the distribution of the reference set plots according to dominant species. A set of treeless (Nfo) reference plots from forest land was added to improve estimates of small wood volume values. The "Nfo" sample plots had only some seed trees detectable from lidar height data. The majority of the reference plots are dominated by Silver birch, Norway spruce or Trembling aspen. The birch and spruce dominated reference plots are distributed over all possible stand ages but there are only few aspen dominated plots in the age range of $20 < A < 50$ (Figure 2). In this study 444 reference plots were used in *k*NN imputations.

For the validation of *k*NN imputations we used data from the network of permanent forest growth research plots (FGN plots) which covers the whole Estonia (Kiviste & Hordo, 2002). The FGN plots in Laeva were established during the period 1995–2004. A circular FGN plots may have a radius of 15, 20, 25 or 30 meters following the rule that at least 100 trees of the first tree layer are contained within the sample plot area. Sample plot radius is increased when this criterion is not fulfilled. On each plot, the polar-coordinates (azimuth and distance from the plot centre), the diameter at breast height, and defects are assessed for each tree. The tree height and height to crown base are measured on every fifth

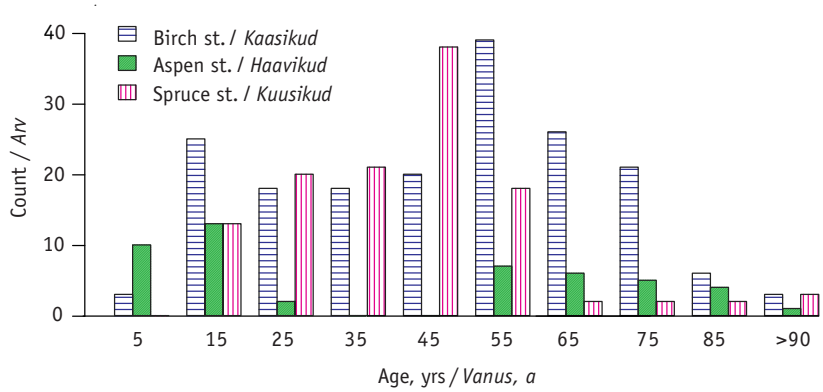


Figure 2. The distribution of training (reference) plots according to forest age and dominant species of the three most widespread species.

Joonis 2. Kolme enamlevinud puuliigi puistute õpetusalade arv vanuse järgi.

Table 1. The reference plot count, average wood volume of the first tree layer (M_1) and average species composition by dominant species.

Tabel 1. Treeningproovitükkide arv, keskmine esimese rinde tüvemaht (M_1) ja liigiline koosseis enamuspuuliigi järgi.

Dominant species / *Enamuspuuliik*	Code / *Tähis*	Plot count / *Proovitükke*	M_1	Average species composition / *Keskmine koosseis*							
				HB	KS	KU	LM	LV	MA	SA	TL
Populus tremula L. / *Haab*	HB	50	262	79	15	3	2	1	0	0	0
Betula pendula Roth / *Kaask*	KS	182	215	8	72	8	7	3	0	0	2
Picea abies (L.) Karst. / *Kuusk*	KU	123	208	2	15	78	1	2	1	0	1
Alnus glutinosa (L.) Gaertn. / *Sanglepp*	LM	19	188	0	23	6	66	4	0	0	1
Alnus incana (L.) Moench / *Hall lepp*	LV	11	68	6	23	2	0	63	0	0	6
Pinus sylvestris L./ *Mänd*	MA	12	293	0	5	13	0	0	82	0	0
Fraxinius excelsior L. / *Saar*	SA	1	278	0	0	22	0	0	0	28	50
Other / *Teised*	TL	3	70	7	17	14	12	8	0	0	42
Treeless / *Lagedad alad*	Nfo	43	-	-	-	-	-	-	-	-	-

tree and also on dominant and rare tree species. The sample plots have measurement interval of five years. The subset of 89 sample plots used in this study was measured on years 2010 and 2011. The wood volume was predicted for the year 2013 based on the increment of the last measurement interval.

The *k*NN predictions were also validated on the stand-wise forest inventory data. The database and digital stand maps were obtained from Estonian State Forest Management Centre. The database had notes about thinnings and field estimates of main inventory variables for each species in each layer (stand elements) for all stands. The

*k*NN predicted stem volume was validated on a subset of 2290 stands which had 1) size over 2 ha, 2) had no treatments or had treatment date earlier than inventory, and 3) were not outliers in the wood volume to spectral reflectance relationship due to disturbances. The wood volume in database was incremented to the year 2013 using algebraic difference model of volume growth (Kangur *et al.*, 2007). This subset was further restricted to 986 stands for the species composition analysis by setting the minimum number of 20 m pixels within stand to n_{pix}> = 37.

Spectral data and lidar data

A cloud free Landsat-8 Operational Land Imager (OLI) image from 03.08.2014 was downloaded from USGS archive and projected to Estonian coordinate system EPSG:3301. Pixel size was set to 20 m for compatibility with lidar metrics feature space. The pixel values in digital numbers (DN) of downloaded image were used in calculations.

The lidar point cloud height statistics – mode, height percentiles, and standard deviation (Table 2) – were calculated from point clouds with 1.3 m height filter to remove near ground reflections in FUSION/LDV (v3.42) software (McGaughey, 2014) by using 20 m cell size for sampling. Filtering was done to reduce the influence of the dense understory vegetation as found out in previous studies (Lang *et al.*, 2012; Arumäe & Lang, 2013). The digital terrain model (DTM) for point cloud height normalization was created using FUSION/LDV modules *Groundfilter* and *GridSurfaceCreate* with cell size set to 4 m.

The sets of feature variables

The characteristic relationships of wood volume (M_{tr}) and Lorey's height of the forest in the reference plots (H_{tr}) with some spectral bands of Landsat-8 OLI image and airborne lidar metrics are shown in Figure (3). Spectral reflectance of the forests decreases with increase in wood volume nonlinearly and in near infrared (OLI5) and shortwave infrared (OLI6) bands the relationship depends on tree species composition (Figure 3 a–c). There are some outliers in the Lorey height to lidar H_{90} relationship (Figure 3 d) which were the training plots with tall seed trees and substantially lower young tree layer of both the layers having similar basal area. Some reference plots with only tall seed trees were also outliers in stem wood volume to lidar H_{90} relationship (Figure 3e) thus, canopy cover (*CaC*) must be included to feature variables for correct estimates of wood volume in such target set pixels.

Many authors emphasize the importance of feature variable selection for *k*NN imputation. We did not have a genetic algorithm interface or bootstrapping interface on the used *k*NN implementation, but constructed five sets of feature variables (Table 2). Two sets consisted of original spectral bands or lidar variables and three sets were based on principal components of the first two sets to reduce dimensionality of feature space and reduce the number of redundant variables Table 1 in McRoberts, 2012. The first set of features consisted of first seven Landsat-8 OLI optical bands and the set is further in the text indicated by L8. The second set ($L8_{PCA}$) of feature variables was the first three principal components of the set L8. The principal components described 60.8%, 36.8% and 1.7% of the total variation in the L8. The third set of feature variables (ALS) was based solely on airborne lidar data and consisted of 90th and 25th percentiles (H_{90} and H_{25} correspondingly) of point cloud height distribution, canopy cover (Lang 2010) at reference height h_{ref} = 3m, and pulse split variable $R1_{ratio}$ = *count(return = 1) / count(all)*. The variables performed well for stem volume estimation in previous study in Aegviidu test site, Estonia (Lang *et al.*, 2012). Fourth set of feature variables (ALS_{PCA}) consisted of the first three principal components of all lidar metrics and the three components described 92.4%, 6.1% and 1.2% of the to-

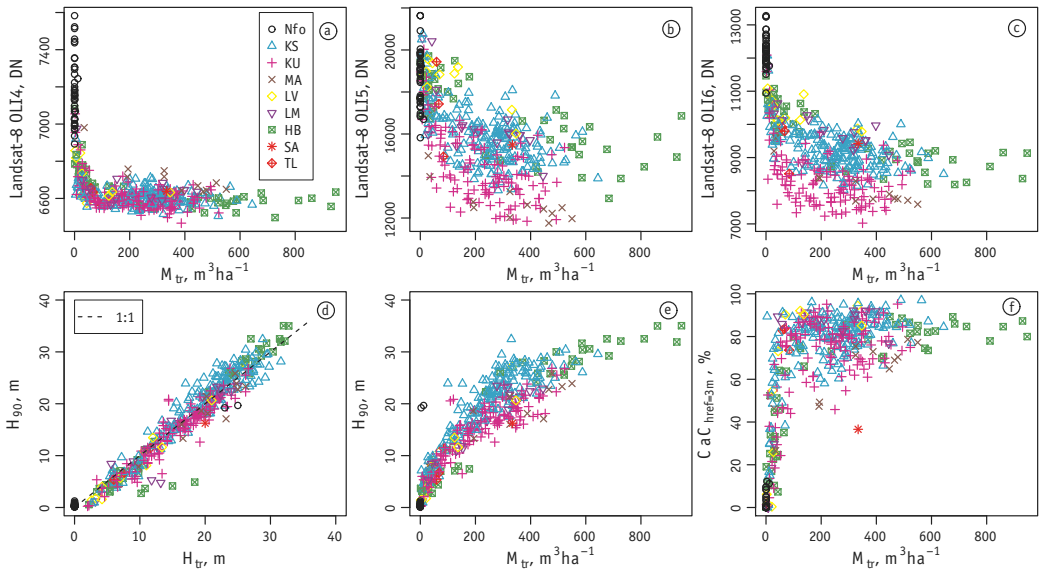


Figure 3. Relationships between some main remote sensing variables and forest inventory variables in the training reference plots. The relationships with wood volume (M_{tr}) are nonlinear, dominant species dependent and often saturating. Species codes are given in Table 1.

Joonis 3. Mõnede kaugseiretunnuste ja metsa takseertunnuste seoseid õpetusaladel. Seosed tüvemahuga (M_{tr}) on mittelineaarsed, liigiomased ja tihti küllastuvad. Lorey kõrguse (Htr) joonisel (d) eristuvad seosest sarnase rinnaspindalaga kõrge ülarinde ja oluliselt madalama esimese rindega proovitükid. Puuliikide koodid on tabelis 1.

tal variation leaving only about 0.3% of the lidar data variability undescribed. The fifth feature variable set (PCA) consisted of the first five principal components of all L8 and lidar variables combined. The five principal components described 93.6%, 4.3%, 1.0%, 0.5% and 0.3% of total variation. IDRISI Taiga (Clark Labs, Worcester, MA, USA) was used to extract principal components.

kNN imputation of wood volume and dominant species

The weighted *k*-nearest neighbour algorithm used in this study was implemented according to method outlined by Franco-Lopez *et al*. (2001). First, Euclidean distance measured in feature space between pixel to be estimated p and pixels with reference data p_i is calculated:

$$d_{p_i,\,p} = \sqrt{\sum_{j=1}^{n_f} a_j^2 (s_{p,\,j} - s_{p_i,\,j})^2}\,, \qquad (1)$$

where $s_{p,\,j}$ = digital number of pixel to be estimated, $s_{p_i,\,j}$ = digital number of pixel with known reference data, n_f = number of dimensions of feature space used and a_j = weight of dimension j. The a_j was fixed to 1 in most of the tests in this study, however, a a_j sensitivity study was carried out on ALS feature variable set.

The reference set pixels are ordered according to their distance from pixel to be estimated so as $d_{p_1,\,p} \le d_{p_2,\,p} \le d_{p_3,\,p}$, ..., $\le d_{p_k,\,p}$.

is probably related to the errors in scaling Using the selected number of nearest neighbours k and previously calculated distances $d_{p_i,\,p}$, the weight of each pixel with known reference data in pixel to be estimated p is calculated:

$$w_{p_i,p} = \frac{1}{d^t_{p_i,p}} \Big/ \sum_{j=1}^{k} \frac{1}{d^t_{p_j,p}}, \qquad (2)$$

where t is reference observation weight. Finally, arbitrary variable m of interest for pixel p is imputed as:

$$\hat{m}_p = \sum_{i=1}^{k} w_{p_i,p} m_{p_i}\ , \qquad (3)$$

where m_{p_i} = attribute value for pixel p_i. For categorical type of variables, mode value is used instead of weighted average. For pixel level error estimation, leave-one-out cross validation method (LOOC) is carried out, where validation data set is constructed so that for estimating a pixel with known attribute data, the point's own data is omitted from the calculations (Katila *et al.*, 2001).

For each target set pixel the wood volume for the first and second tree layer ($M1_{kNN}$, $M2_{kNN}$), and dominant species code was predicted by using the five sets of feature variables (L8, $L8_{PCA}$, ALS, ALS_{PCA}, PCA). The number of tested nearest neighbours was $k \in (1, 2, 3)$. The sensitivity of the *k*NN imputed wood volume to feature-based weight a_j was carried out on ALS feature variable set by assigning the a_j values given in Table 3. In the test setup values $k = 3$ and $t = 2$ were used. The sensitivity of the *k*NN imputed wood volume to distance-based neighbour weight t (Eq. 2) was tested on feature variable sets L8, ALS and PCA by using values of $t \in (1, 2, 3)$.

Table 2. The variables in feature variable sets. The principal component feature variable sets were based on the included variables. H_x are the lidar point cloud height distribution percentiles. CaC_x are the lidar data based canopy cover estimates at reference height shown in subscript. *Stdev* is standard deviation and *Mode* is the mode value. $R1_{ratio}$ is calculated from point cloud as count(return=1)/count(all). OLI corresponds for Landsat-8 Operational Land Imager.

Tabel 2. Kokkuvõte kNN ennustustes kasutatud tunnuste komplektidest. Peakomponentidest koosnevate tunnuste komplektid põhinevad näidatud tunnustele. H_x on punktiparve kõrgusjaotuse vastavad protsentiilid. CaC_x on punktiparvest alaindeksis näidatud referentskõrgusel arvutatud katvus. Stdev ja Mode tähendavad standardhälbe ja moodväärtuse arvutamist. $R1_{ratio}$ arvutatakse punktiparvest seosega arv(peegeldus=1)/peegeldusi. OLI tähendab Landsat-8 Operational Land Imager vastava numbriga spektraalkanaleid.

Variable / *Tunnus*	The feature variables / *Kirjeldavate tunnuste komplektid*				
	L8	ALS	$L8_{PCA}$	ALS_{PCA}	PCA
H_{10}..H_{80}, H_{99}	0	0	0	1	1
H_{25}	0	1	0	1	1
H_{90}	0	1	0	1	1
$CaC_{1.3m}$	0	0	0	1	1
$CaC_{2.0m}$	0	0	0	1	1
$CaC_{3.0m}$	0	1	0	1	1
$Stdev(H_{ALS})$	0	0	0	1	1
$Mode(H_{ALS})$	0	0	0	1	1
$R1_{ratio}$	0	1	0	0	0
OLI1..OLI7	1	0	1	0	1

A special test was carried out with feature variable sets L8, ALS and PCA with the number of birch dominated reference plots substantially decreased. In the forest age range $10 \le A < 20$ 50% of the available reference plots were included while in the forest age range $20 \le A < 70$ only 30% of the available plots were included by random selection. The number of neighbours in the special test was $k = 3$.

Validation of the predictions

A simple method for validation of the *k*NN imputation results is the leave-one-out cross validation where the validation statistics are calculated on the imputed values for the reference plots. The plot for which the values are imputed is always excluded from the reference set. However, Tomppo *et al*. (2009) warn that LOOC does not always produce realistic error estimates. Therefore, in addition to LOOC we carried out a validation on two independent datasets: 1) 89 forest growth network plots (FGN plots), and 2) stand-wise forest inventory database (IDB records).

The imputed wood volume for FGN plots was extracted from raster by using the areas of intersections of pixels (squares) and FGN plot geometry (circle) as weights on pixel values (Lang *et al*., 2005). This approach was preferred over the whole pixel extraction, since the radius of the FGN plots ranged from 20 m to 30 m and pixel size was 20 m. For each forest stand the pixels from *k*NN imputed wood volume maps were extracted according to the pixel centre location. For each IDB stand the average wood volume including both the first and second layer over all extracted pixels was calculated. The IDB wood volume is known to be underestimated (Raudsaar *et al*., 2014) compared to sample plot based estimates e.g. National Forest Inventories (NFI). Therefore the IDB wood volume was scaled by using linear regression on wood volume (M_{tr}) of the reference plots (Figure 4).

The imputed dominant species composition was validated only on IDB stands which were sufficiently large. The species composition for each IDB stand was calculated from the imputed dominant species map based on the relative share of pixels with particular tree species codes. So, for a stand with 50% spruce and 50% birch the correct number of pixels with main species code corresponding to spruce was expected to be 50% of all pixels within the stand polygon.

Root mean square error (*RMSE*) was calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \quad (4)$$

where y_i = attribute value, $\hat{y}_i$ = attribute value estimated by kNN method, n = number of observations. Bias was estimated as:

$$\bar{e} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}. \quad (5)$$

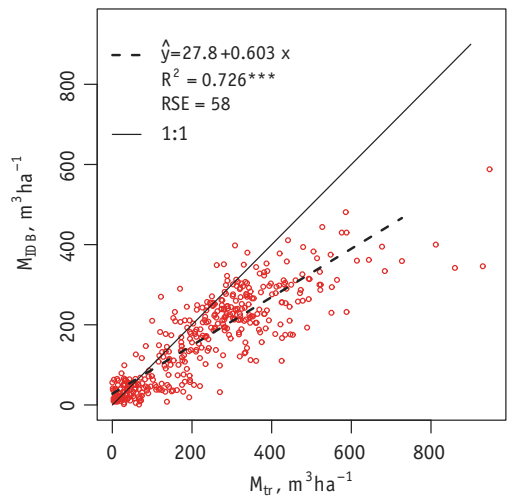


Figure 4. The linear regression between total wood volume given in the stand wise forest inventory database (M_{IDB}) and the wood volume (M_{tr}) from corresponding reference plots was used to scale the M_{IDB} to M_{tr} range. The RSE is the model residual error given by lm procedure in R software (R Core Team, 2015).

Joonis 4. Takseerandmebaasis olev puistu tüvemaht M_{IDB} teisendati lineaarse regressioonimudeliga õpetusproovitükkidel mõõdetud tüvemahu M_{tr} skaalasse. RSE on tarkvara R (R Core Team, 2015) protseduuriga lm lähendatud mudeli jääkhälve.

Results

Here we present only results from $k = 3$ imputations based on L8, ALS and PCA feature datasets since there were only small differences in the results if k was changed or $L8_{PCA}$ was used instead of L8 or ALS_{PCA} was used instead of ALS feature variable set. The decrease of RMSE and overall bias estimate obtained from LOOC validation of imputed wood volume showed that inclusion of airborne lidar variables did substantially improve the predictions for the dominant tree layer (Figure 5 a–c). Similar results are obtained by other authors. The PCA feature variable set performed best and most of the increase of estimation accuracy was gained on aspen stands which had the largest wood volume compared to other forests. The predicted wood volume for the second layer of trees had substantial scatter independent from the feature variable set (Figure 5 d–f). The well-known systematic overestimation of small volumes and systematic underestimation of large wood volumes was also present in our imputation results. This lack of fit has been assumed to be a linear function of wood volume (Holmgren *et al.*, 2000); however, in our results the lack of fit for the upper tree layer wood volume first starts to increase at small values, then reaches its positive maximum and then starts constantly to decrease crossing the zero value near to the mean of measured wood volume.

The imputed wood volume validation on forest growth network plot data produced similar results to LOOC (Figure 6). The overall lack of fit for the first tree layer wood volume was negligible when using

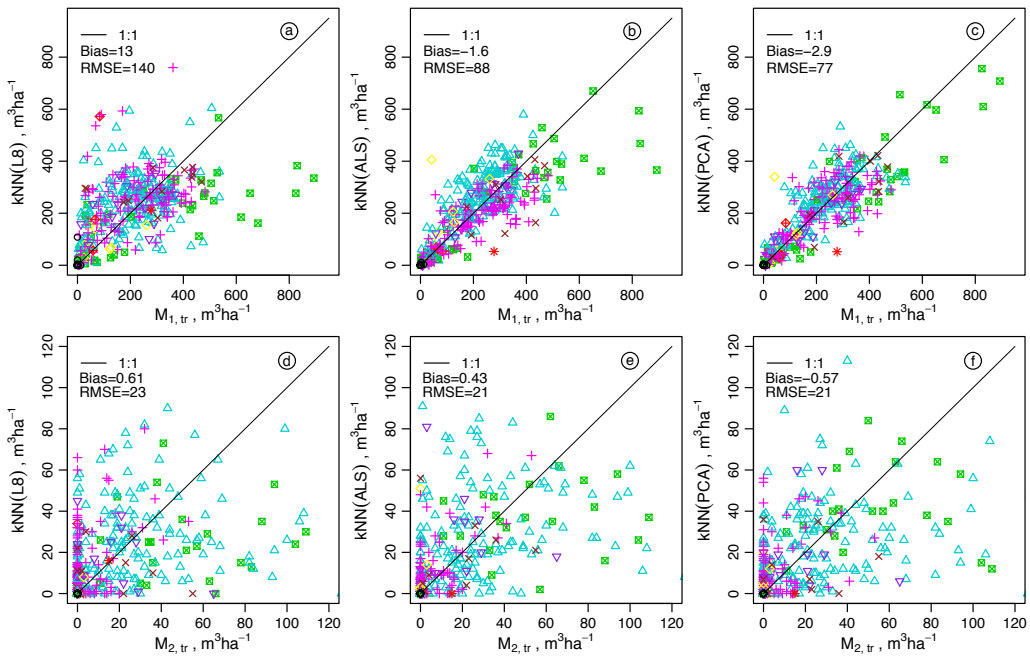


Figure 5. The leave-one-out cross validation of imputed wood volume for the first layer (M_1) and second layer (M_2) of trees (in rows) using three feature datasets (in columns). The symbols are explained in Figure 3.

Joonis 5. Õpetusalade andmetel jäta-üks-välja meetodil tehtud kNN ennustuste valideerimise tulemused. Ridades on eraldi esimese- ja teise rinde tüvemahtud (M_1, M_2) ja tulpades on aluseks olnud erinevad andmekomplektid. Sümbolid on samad mis joonisel 3.

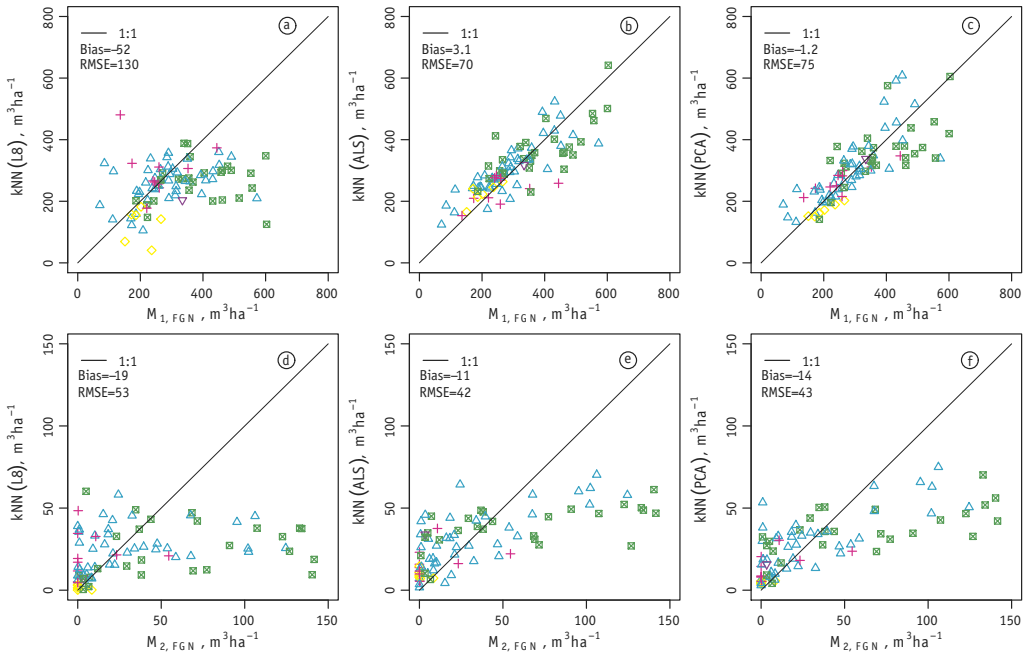


Figure 6. The validation of imputed wood volume for the first layer (M_1) and second layer (M_2) of trees (in rows) using three feature datasets (in columns) on forest growth plots. The symbols are explained in Figure 3.

Joonis 6. kNN tüvemahu valideerimise tulemused metsa kasvukäigu võrgustiku proovitükkidel. Ridades on eraldi esimese ja teise rinde tüvemahtud (M_1, M_2) ja tulpades on aluseks olnud erinevad andmekomplektid. Sümbolid on samad mis joonisel 3.

ALS or PCA feature variable sets. Since the small and large wood volumes were not present in the FGN plots, the systematic overestimation or underestimation was not apparent. The difference in results based on ALS or PCA feature variable sets was small according to RMSE and overall bias and was probably influenced mainly by random errors inherent in georeferencing and other technical aspects in data processing chain. The predicted volume of wood for second layer had much less scatter compared to that found in the reference set LOOC. The reason for more stable estimates may be related to bigger size of FGN plots which cover several 20 m pixels whereas single pixel values were used for the reference set LOOC in our *k*NN implementation. However, larger wood volume values of the second tree layer were systematically underestimated (Figure 6 d–f).

In validation of imputed wood volume against forest inventory database records, the PCA feature variable dataset, which was based on all possible remote sensing variables, ranked the best (Figure 7). Considering the mean total wood volume 134.6 m^3 ha^{-1} of the validation forest stands the relative RMSE for the PCA feature variable dataset based wood volume estimation was 44%. The second best was ALS feature variable set based wood volume estimate with larger bias (18.9 m^3 ha^{-1}) and RMSE (61.9 m^3 ha^{-1}) while the only spectral information (L8) based estimate had the largest bias (20.6 m^3 ha^{-1}) and RMSE (89.9 m^3 ha^{-1}). All predictions had overestimation at small and underestimation at large

wood volumes. The overall positive bias of sample plot based *k*NN estimates into stand-wise estimates (Figure 4).

In feature-based weight a_j influence test the smallest RMSE of wood volume estimates according to LOOC was obtained at a_j for canopy cover $CaC_{3.0m}$ and weights of other ALS variables fixed to values given in Table 3. The optimal a_j apparently scaled the estimates of canopy cover and canopy height 90th percentile into the same range. Either the increasing or decreasing of a_j made estimation errors larger. Validation of the wood volume estimates on IDB data however, showed a slight increase in bias (24.7 m^3 ha^{-1}) and a marginal decrease in RMSE (60.1 m^3 ha^{-1}) compared to unweight case.

Table 3. The feature weighting parameter a_j (Eq. 1) values which were used in the sensitivity study on ALS feature variable set. The ALS variables values were scaled to integer numbers to decrease required data storage space. The values were input into (Eq. 1) and further modified by a_j.

Tabel 3. Kirjeldavate tunnuste kaalu a_j *(1) mõju tüvemahu ennustusele uuriti ALS andmekomplektil. ALS tunnuste väärtused skaleeriti täisarvudeks andmemahu kokkuhoiuks. Neid väärtusi kasutati sisendina valemis (1) ja muudeti edasi kaaluga* a_j.

Feature variable / *Kirjeldav tunnus*	Range / *Haare*	a_j
$CaC_{3.0m}$	0..10000	0.01..1.0
H_{90}	0..3769	1.0
H_{25}	0..2791	0.2
$R1_{ratio}$	0..100	1.0

The change in distance-based neighbour weight *t* value from 0 to 3 resulted only in a small, but still detectable improvements in estimation accuracy as found for forest stands. The bias estimate and RMSE of wood volume decreased about 1.5 m^3 ha^{-1} when *t* was increased from zero to three. Similar trend was detectable by validation of estimates on FGN plots, except for first tree layer wood volume estimate based on L8 feature variable set.

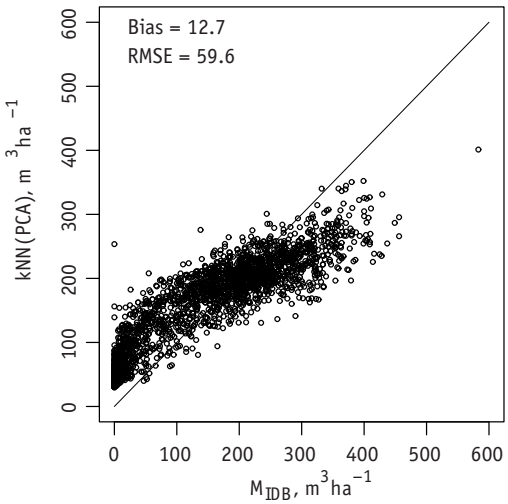


Figure 7. The validation of PCA feature variable set based wood volume (M_1+M_2) estimate on stand-wise forest inventory data.

Joonis 7. Peakomponentide andmestikul (PCA) põhineva tüvemahu (M_1+M_2) hinnangu valideerimise tulemused metsakorralduse andmetel.

The second important response variable was the dominant species code. In Figure 8 the dependence of the predicted proportions of birch, spruce and aspen are shown based on feature variable sets L8, ALS and PCA. The overall predicted proportion of birch had about 10% positive overall bias (Figure 8 a, d, g). The proportion of birch trees was underestimated in young stands and overestimated in old stands, regardless of the feature variable set used. The proportion of spruce trees was underestimated, being the smallest for the ALS feature variable set at the price of much larger random errors. Visual inspection of the maps revealed substantial noise in the ALS feature variable set based dominant species maps. The predicted proportion of spruce trees had also notable but less pronounced discrepancy

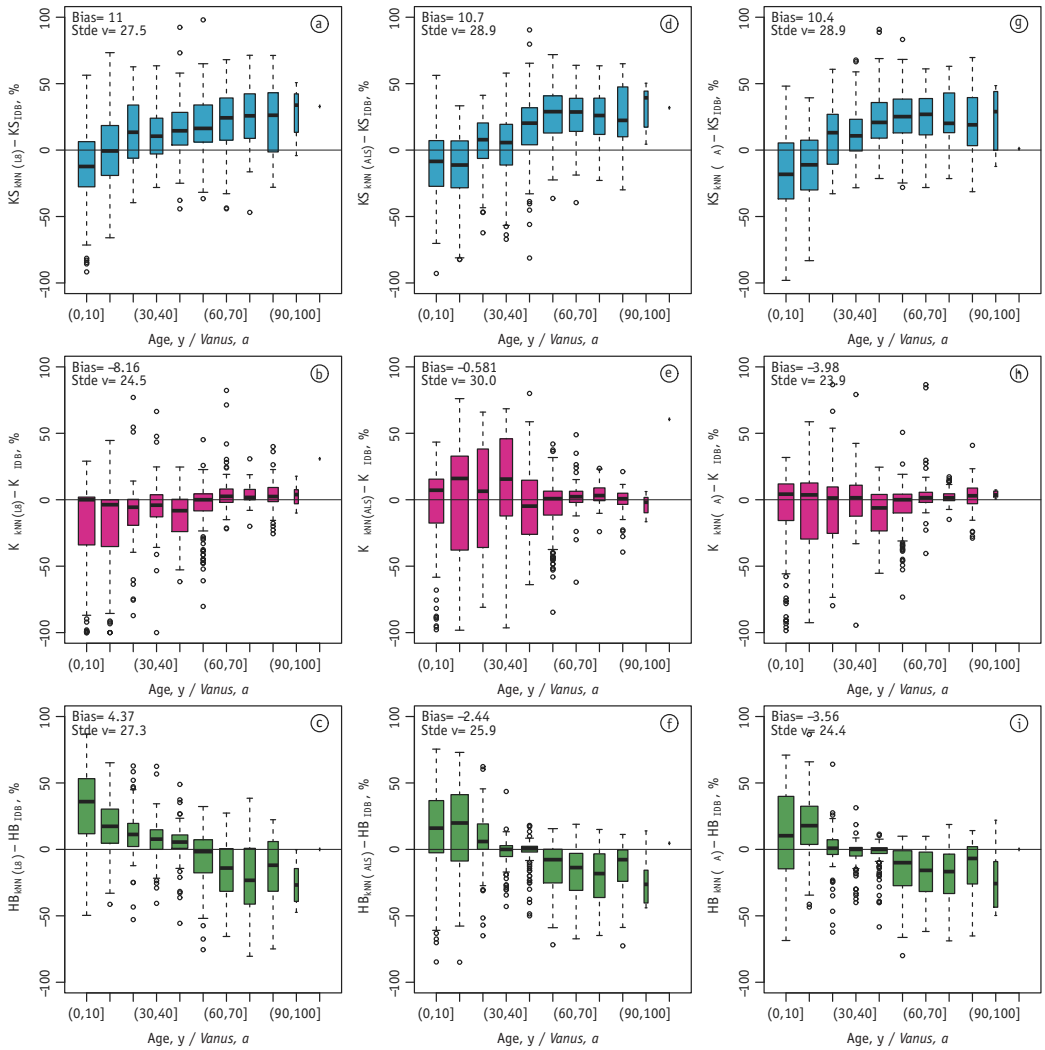


Figure 8. The difference of species composition between stand-wise forest inventory data and *k*NN imputed value in age classes using three different feature datasets (columns). Stdev is the standard deviation of the difference between predicted proportion and the measured value.

Joonis 8. Metsaeraldistele kNN abil ennustatud kase, kuuse ja haava osakaalude võrdlus takseeritud osakaaludega vanusrühmade kaupa. Stdev on ennustatud osakaalu ja andmebaasis antud osakaalu erinevuse standardhälve.

from observed value depending on forest age (Figure 8 b, e, h). The difference of predicted and observed share of aspen trees, however, showed exactly the opposite trend compared to the prediction error of birch proportion. The smallest discrepancy of aspen proportion was in the age classes in which aspen dominated plots were not present in the reference set. In the same age classes also the predicted proportion of aspen was the smallest (Figure 8 c, f, i). The comparison of species composition in the reference plots with the corresponding forest stand records in the inventory database did not reveal such forest age dependent lack of fit (Figure 9). Hence, the cause

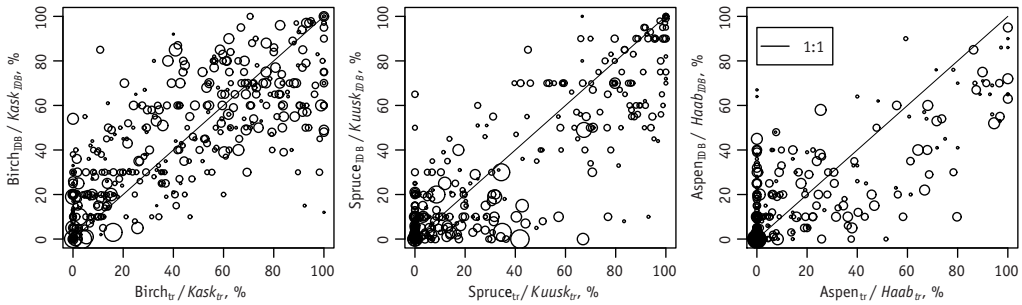


Figure 9. The comparison of proportions of three most widespread tree species on forest inventory stands and corresponding reference set plots. The marker size is dependent on forest age. There are no clusters and no trend detectable in respect to forest age.

Joonis 9. Kase, kuuse ja haava osakaalu võrdlus kNN treeningproovitükkidel ja neid sisaldavatel eraldistel. Sümbolid on vanuse järgi skaleeritud. Jooniselt ei ilmne vanusest sõltuvaid klastreid ega trendi.

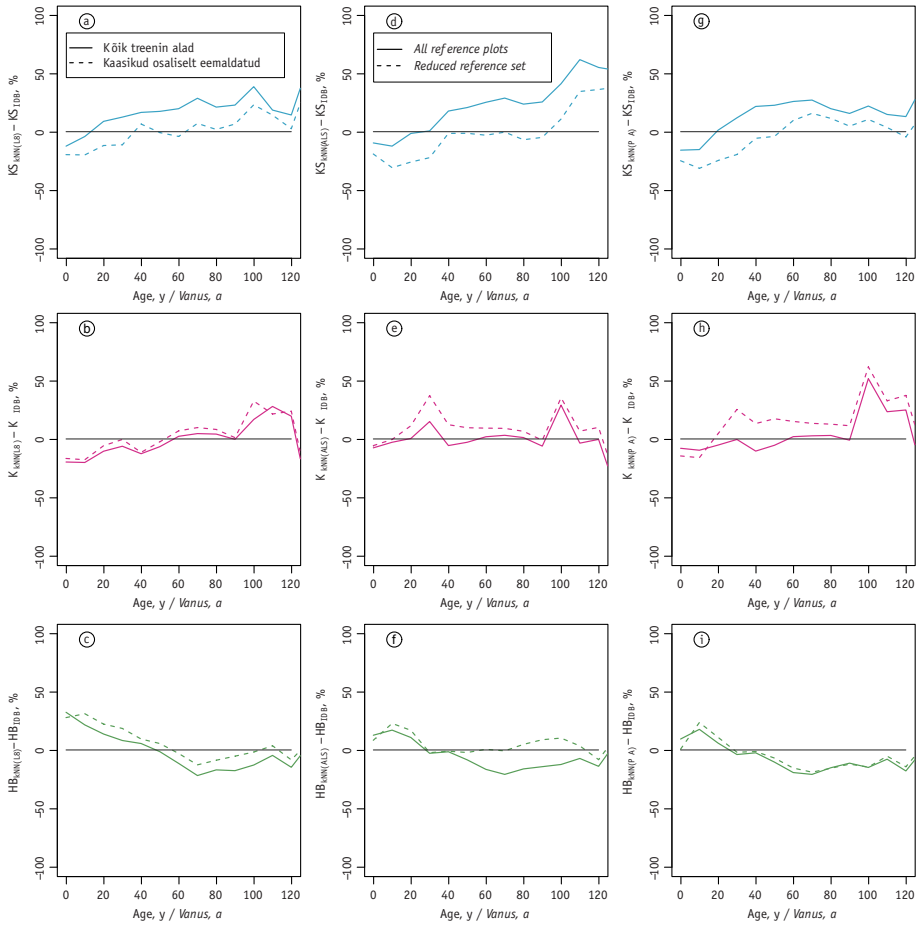


Figure 10. The effect of partial removal of birch dominated plots from the reference set on predicted proportions of birch, spruce and pine in case of three different feature datasets (columns).

Joonis 10. Kaseenamusega treeningalade osaline eelmaldamine õpetusalade hulgast avaldab mõju enamuspuuliigi ennustusele pikslitel ja seeläbi ka puistute liigilise koosseisu hinnangutele.

is rather in the *k*NN imputation procedure than a possible systematic error in the forest inventory database.

The birch dominated reference plot removal decreased also the predicted proportion of birch dominated pixels but did not always correct for the previously described opposite to birch proportion trend in the share of aspen (Figure 10). The partial removal of birch dominated reference plots caused also the predicted share of spruce to increase. We repeated the birch plot partial removal test with random selection several times and always reached the same general changes in predicted species composition. This can be partially explained by the fact that many reference plots were mixed stands and the spruce dominated plots had also birch and other deciduous trees, which had an influence on the feature variables too. The results also expose the dependence of nearest neighbour technique on the distribution of the reference plots in respect to response variables and feature variables. An explanation here is that the probability of a reference observation to be used as nearest neighbour in imputation increases with the count of such plots which have similar vectors of feature variable values. Similar note is given in the manual of *k*NN implementation in IDRISI Taiga (Clark Labs, Worcester, MA, USA).

Discussion

In this study we were not able to repeat the *k*NN imputation of wood volume for forest stands with accuracy reported by Packalén & Maltamo (2007) concerning the estimation bias at small or large predicted values. Packalén & Maltamo (2007) carried out their tests in boreal forests which were dominated by Norway spruce and Scots pine and the share of deciduous was in average only about 10%. Another reasons for the good accuracy statistics achieved in the study of Packalén & Maltamo (2007) who report 10% relative RMSE for average stem volume at stand level, are the stand level leave-one-out cross validation which tends to produce more optimistic figures (Tomppo *et al*., 2009), and the use of airborne images as spectral feature variables which provide much finer spatial resolution compared to Landsat-8 OLI images used in this study. Concerning the feature variable selection, we can confirm that already a few informative variables e.g. principal components of original feature variable set perform sufficiently well as pointed out by Packalén & Maltamo (2007) for canonical components used in *k*MSN.

The inclusion of airborne lidar metrics did improve the accuracy of *k*NN predicted wood volume compared to spectral feature variables but did not remove the underestimation of the wood volumes at larger values for the lower layer of trees. Neither did the ALS feature variable set nor PCA feature variable set, which both included information from lidar metrics, entirely remove the well-known (Fazakas *et al*., 1999) overestimation of small values and underestimation of large values in the predicted wood volume in the upper tree layer. While airborne lidar data are still quite expensive for large area applications, the medium spatial resolution multi spectral satellite images still remain one of the main feature variable sources. The spectral information from airborne or space borne measurements must be used in addition to lidar data if tree species composition is target in *k*NN imputation. Tree species composition can be predicted in some extent based on lidar data only (e.g. Breidenbach *et al*., 2010), since the airborne topographic lidar emit pulses in the near infrared part of electromagnetic spectrum where the spectral reflectance of coniferous and deciduous forests is different (Nilson & Peterson, 1993**)**. The different spectral reflectance impacts the formulation of the pulse reflection positions and the relationships between lidar metrics and forest height or wood volume will be species dependent. Recent studies (Zald *et al*., 2014) show that

time series of Landsat-5 TM and Landsat-7 ETM+ type satellite images may further improve the accuracy of species composition prediction. Strategic forest inventory data and multitemporal Landsat images can also be used for basal area prediction and the estimates may provide an alternative to expensive sample plot based forest management inventories (McRoberts, 2008). However, in practical applications a careful validation of the imputed maps of forest inventory variables is required to decide if the accuracy of the maps is suitable for decision making.

The accuracy estimation of the *k*NN imputed maps is not a trivial task and may be related to many difficulties. The simplest method is the leave-one-out cross validation on the reference set but this may produce unrealistically optimistic results (Tomppo *et al*., 2009). An alternative is to split the reference set into training and validation sets which will have a negative impact on the final estimation accuracy (Chirici *et al*., 2008). McRoberts (2012) proposes bootstrapping with replacement to construct a large number alternative reference sets based on the original reference set. For each bootstrap sample the nearest neighbour technique is to be then used to calculate predictions for each population unit. Variability and bias will be then estimated using the bootstrap samples based predictions (McRoberts, 2012). In this study we conducted a test where large proportion of birch dominated reference plots was removed to equalize forest reference plot distribution of different deciduous forests according to forest age. This treatment fulfilled partially its purpose and the overestimation of birch proportion decreased in forest stands, however, it also caused positive bias in the spruce proportion which was predicted unbiased for most of the age classes in case of using full reference set. This experience indicates that if target set forests in an area of interest, e.g. within a Landsat image frame, have systematically different characteristics by regions then a bootstrap sample dominated by the reference set units from one region may produce unfavourable estimates for the other region. A solution could be the application of *k*NN techniques and bootstrapping regionally based on some stratification criterion in similar to Tomppo *et al*. (2009) who used separate reference sets for mineral soils and peat land soils.

The *k*NN procedure does not require any assumptions regarding the distributions of response or predictor variables, instead, user has to select the best performing feature variables, make the choice of distance metric, and find appropriate values for k, t, and a_j. In our study the selection of feature variables (mainly inclusion of airborne lidar metrics) had the largest impact on results while varying the number of neighbours or distance-based neighbour weight t or feature-based weight a_j had much smaller influence. By using fewer neighbours the lack of fit close to the minimum and maximum values to be estimated decreased in some extent but was still apparent. The increase in t from 0 to 3 decreased both RMSE and bias estimate at forest stand level by about 1.5 m^3 ha^{-1} independent from feature variable set. On the FGN plots the improvement was up to 5 m^3 ha^{-1} for PCA feature variable set, but smaller for other feature variable sets. The problem of selecting appropriate values k, t, and a_j and the best set of features for a regional *k*NN application can probably be solved by adding an optimization routine to *k*NN implementation similar to McRoberts (2008, 2012) and McRoberts *et al*. (2015). However, the optimization may have to be carried out separately for wood volume estimation and for species composition estimation, since wood volume is related to forest height and canopy cover, but species separation is better based on spectral reflectance. Inclusion of spectral feature variables to ALS feature variable set may increase redundancy and wood volume estimation errors. On the other hand, if species composition is calculated

from shares of the tree species wood volume in target set sampling units then the optimization routine can be adopted for both by keeping attention also to the other important measures e.g. species composition in different age classes.

Optimization may tune *k*NN for a certain region, for a particular set of features or for a set of response variables based on validation observations exclusive to reference set. However, such exclusive validation datasets are not always available or require additional field measurements increasing the project cost. In this study stand-wise forest inventory data and forest growth network sample plots from Laeva test site were used as the exclusive validation information. Both the validation datasets had their positive and negative aspects. Stand-wise forest inventory covers most of the forests in test site, but the inventory variables have bigger random and systematic errors due to the applied inventory method. The FGN sample plots are instrumentally measured, but the count is small and the dataset lacks observations from young stands and there are only few sample plots from spruce stands. Both datasets required updating to predict stem volume information to the reference set plots measurement time. Hence, neither of the validation datasets was perfect to study the causes of under and over-estimation of wood volume. The solution for such general study can be an artificial dataset with controlled shape of relationships between response and feature variables, adjustable distribution of reference observations in respect to response variables and simulated noise. Forest reflectance models and empirical models may be used to create such datasets for *k*NN studies.

Conclusions

Our study confirmed that in forests similar to Laeva test site the *k* nearest neighbour imputed wood volume may be overestimated at small values and underestimated at large values. This lack of fit was present for the wood volume in upper and lower tree layer. Inclusion of airborne lidar data did decrease but not remove this lack of fit. We also found that a few principal components of the original feature variable set already contain enough information for *k*NN imputation.

There was no substantial difference in results when using principal components instead of original feature variables, which shows the principal component analysis as an efficient tool for feature space dimensionality reduction. The *k*NN predicted species composition had a forest age dependent lack of fit at species level and the discrepancy was dependent also on the number of reference plots with similar dominant species. This finding can be important for next studies which use bootstrap samples with replacement for error estimation and for confidence limits construction of the *k*NN imputed values over study areas consisting of regionally different forests. Considering the very basic and not optimized *k*NN implementation used in this study the results are still encouraging, since the forest in Laeva test site are far more complex than used in many other studies.

Acknowledgements. The authors thank Professor Andres Kiviste and Dr. Jan Pisek for commenting of the manuscript. Comments from anonymous reviewers helped substantially improve the manuscript. The establishment of the long-term forest growth monitoring plots was supported by the Estonian Environmental Investment Centre. The authors would like to thank the Estonian Land Board for the airborne lidar data and USGS for the Landsat-8 OLI image. Data acquisition was financed by Estonian State Forest Management Centre. Data analysis was supported by Estonian Research Council grants SF0180009Bs11, IUT21-4 and by the European Regional Development Fund (CECT) project ERMAS.

References

Adermann, V. 2010. Development of Estonian National Forest Inventory. – Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E. (eds.). National Forest Inventories. Heidelberg, Springer, 171–184.

Arumäe, T., Lang, M. 2013. A simple model to estimate forest canopy base height from airborne lidar data. – Forestry Studies / Metsanduslikud Uurimused, 58, 46–56.

Breidenbach, J., Nothdurfth, A., Kändler, G. 2010. Comparison of nearest neighbour approaches for small area estimation of tree species-specific forest inventory attributes in central Europe using airborne laser scanner data. – European Journal of Forest Research, 129, 833–846.

Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D., Maselli, F., Bertini, R. 2008. Non-parametric and parametric methods using satellite images for estimating growing stock volume in alpine and Mediterranean forest ecosystems. – Remote Sensing of Environment, 112, 2686–2700.

Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., Koch, B. 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. – Remote Sensing of Environment, 154, 102–114.

Fazakas, Z., Nilsson, M., Olsson, H. 1999. Regional forest biomass and wood volume estimation using satellite data and ancillary data. – Agricultural and Forest Meteorology, 98/99, 417–425.

Franco-Lopez, H., Ek, A., Bauer, M. 2001. Estimation and mapping of forest sand density, volume, and cover type using the k-nearest neighbors method. – Remote Sensing of Environment, 77, 251–274.

Gilichinsky, M., Heiskanen, J., Barth, A., Wallerman, J., Egberth, M., Nilsson, M. 2012. Histogram matching for the calibration of *k*NN stem volume estimates. – International Journal of Remote Sensing, 33, 7117–7131.

Holmgren, J., Joyce, S., Nilsson, M., Olsson, H. 2000. Estimating stem volume and basal area in forest compartments by combining satellite image data with field data. – Scandinavian Journal of Forest Research, 15, 103–111.

Howard, J.A. 1991, Remote sensing of forest resources. Chapman & Hall, London, UK. 420pp.

Huiyan, G., Limin, D., Gang, W., Dong, X., Shunzhong, W., Hui, W. 2006. Estimation of forest volumes by integrating Landsat TM imagery and forest inventory data. – Science in China: Series E Technological Sciences, 49, 54–62.

Kajisa, T., Murakami, T., Mizoue, N., Kitahara, F., Yoshida, S. 2008. Estimation of stand volume using *k*-nearest neighbors method in Kyushu, Japan. – Journal of Forest Research, 13, 249–254.

Kangur, A., Sims, A., Jõgiste, K., Kiviste, A., Korjus, H., von Gadow, K. 2007. Comparative modeling of stand development in Scots pine dominated forests in Estonia. – Forest Ecology and Management, 250, 109–118.

Katila, M., Tomppo, E., 2001. Selecting estimation parameters for the Finnish Multisource National Forest Inventory. – Remote Sensing of Environment, 76, 16–32.

Kiviste, A., Hordo, M. 2002. Network of permanent forest growth plots in Estonia. – Metsanduslikud Uurimused / Metsanduslikud Uurimused, 37, 43–58.

Krigul, T. 1972. Forest mensuration. (Metsatakseerimine). Valgus, Tallinn. 358 pp. (In Estonian).

Lang, M. 2010. Estimation of crown and canopy cover from airborne lidar data. – Forestry Studies / Metsanduslikud Uurimused, 52, 5–17.

Lang, M. Lükk, T., Rähn, A., Sims, A. 2005. Change detection on permanent forest growth sample plots using satellite images. – Forestry Studies / Metsanduslikud Uurimused, 43, 24–37.

Lang, M., Arumäe, T., Anniste, J. 2012. Estimation of main forest inventory variables from spectral and airborne lidar data in Aegviidu test site, Estonia. – Forestry Studies/ Metsanduslikud Uurimused, 56, 27–41.

Latifi, H., Koch, B. 2012. Evaluation of most similar neighbour and random forest methods for imputing forest inventory variables using data from target and auxiliary stands. – International Journal of Remote Sensing, 33, 6668–6694.

Lõhmus, E. 2004. Estonian forest site types. (Eesti metsakasvukohatüübid). Eesti loodusfoto, Tartu. 80 pp. (In Estonian).

McGaughey, R.J. 2014. FUSION/LDV: Software for LIDAR Data Analysis and visualization. March 2014 – FUSION, Version 3.42. United States Department of Agriculture Forest Service Pacific Northwest Research Station.

McInerney, D.O., Nieuwenhuis, M. 2009. A comparative analysis of *k*NN and decision tree methods for the Irish National Forest Inventory. – International Journal of Remote Sensing, 30, 4937–4955.

McRoberts, R.E. 2008. Using satellite imagery and the k-nearest neighbors technique as a bridge between strategic and management forest inventories. – Remote Sensing of Environment, 112, 2212–2221.

McRoberts, R.E. 2012. Estimating forest attribute parameters for small areas using nearest neighbour techniques. – Forest Ecology and Management, 272, 3–12.

McRoberts, R.E., Tomppo, E.O. 2007. Remote sensing support for national forest inventories. – Remote Sensing of Environment, 110, 412–419.

McRoberts, R.E., 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.

Moeur, M., Stage, A. R. 1995. Most similar Neighbor: An improved sampling inference procedure for

natural resource planning. – Forest Science, 41, 337–359.

Nilson, T., Peterson, U. 1994. Age dependence of forest reflectance – analysis of main driving factors. – Remote Sensing of Environment, 48, 319–331.

Packalén, P., Maltamo, M., 2007. The k-MSN method for the prediction of species specific stand attributes using airborne laser scanning and aerial photographs. – Remote Sensing of Environment, 109, 328–341.

Poso, S., Karlsson, M., Pekkonen, T., Härmä, P. 1990. A system for combining data from remote sensing, maps and field measurement for forest planning purposes. – University of Helsinki, Department of Forest Mensuration and Management. Research notes, 23, 40 pp.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. [WWW document]. – URL http://www.R-project.org [Accessed 10 October 2014].

Raudsaar, M., Pärt, E., Adermann, V. 2014. Review of Estonian forest resources. Yearbook Forest 2013. Compiled by Keskkonnaagentuur. OÜ Paar, Tartu, p. 1–2. (In Estonian).

Tamm, T., Remm, K. 2009. Estimating the parameters of forest inventory using machine learning and the reduction of remote sensing features. – International Journal of Applied Earth Observation and Geoinformation, 11, 290–297.

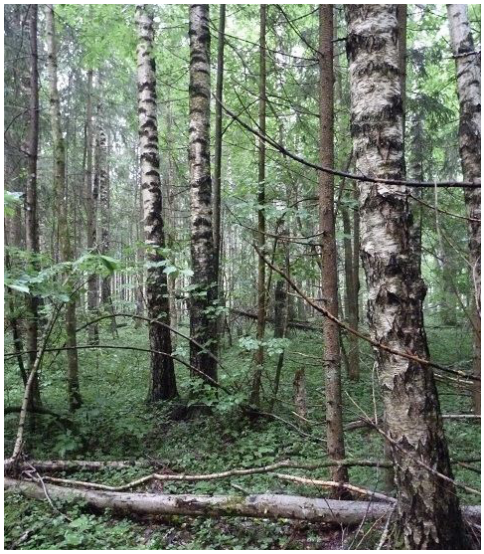
Tomppo, E., Gagliano, C., De Natale, F., Katila, M., McRoberts, R.E. 2009. Predicting categorical forest variables using an improved k-Nearest Neighbour estimator and Landsat imagery. – Remote Sensing of Environment, 113, 500–517.

Tomppo, E., Schadauer, K., McRoberts, R. E., Gschwantner, T., Gabler, K., Ståhl, G. 2010. History of NFIs. – Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E. (eds.). National Forest Inventories. Heidelberg, Springer, 1–2.

Zald, H.S.J., Ohmann, J.L., Roberts, H.M., Gregory, M.J., Henderson, E.B., McGaughey, R.J., Braaten, J. 2014. Influence of lidar, Landsat imagery, disturbance history, plot location accuracy, and plot size on accuracy of imputation maps of forest composition and structure. Remote Sensing of Environment, 143, 26–38.

Appendix 1. Some examples of plots from different forest types in Laeva test site (photos M. Merenäkk).

Lisa 1. Näited erinevatest naadi (a, c, d, f) ja angervaksa (b, e) kasvukohatüübi proovitükkidest Laeva testalalt (fotod M. Merenäkk).



a) 55 years old AG site type birch stand with second layer spruces.

b) 30 years old FP site type birch dominated stand with 20% grey alder in the first layer.

c) 50 years old AG site type birch and spruce mixed forest stand.

d) 70 years old AG site type birch stand with dense understory vegetation.

e) 55 years old FP site type birch stand with dense understory vegetation.

f) 45 years old AG site type spruce dominated stand with sparse second layer spruce.

Puistute liigilise koosseisu ja tüvemahu hindamine k-lähima naabri meetodil mitmerindelistes majandatavates segametsades

Mait Lang, Tauri Arumäe, Tõnu Lükk ja Allan Sims

Kokkuvõte

Jätkusuutlik metsamajandus ja seda toetav metsapoliitika peab tuginema objektiivsetele ja piisavalt sagedasti uuendatud eelistatult kogu käsitletavat ala katvatele takseerandmetele. Üksikvaatlusteks on kas puistud või proovitükid, mille põhjal saab soovi korral teha üldistusi piirkonniti (Krigul, 1972). Suurtel aladel takseerandmestiku uuendamiseks on paljudes maades kasutamist leidnud niinimetatud *k* lähima naabri klassifitseerimismeetod (*k*NN), mis seob puistu takseerimisel või proovitüki mõõtmisel saadud andmed satelliitidelt või lennukitelt tehtud spektraalsete mõõtmiste või kolmemõõtmeliste punktiparvedega võimaldades niimoodi koostada ülepinnalisi takseertunnuste kaarte (Fazakas *et al*., 1999, Holmgren *et al*., 2000; Huiyan *et al*., 2006; McRoberts & Tomppo, 2007; Packalén & Maltamo, 2007; Chirici *et al*., 2008; Kajisa *et al*., 2008; McInerney & Nieuwenhuis, 2009; Breidenbach *et al*., 2010; McRoberts, 2012; Fassnacht *et al*., 2014; Zald *et al*., 2014). Eestis on varem *k*NN meetodit kasutanud masinõppe ühe osana metsade takseertunnuste ennustamiseks lausmetsakorralduse andmete järgi Tamm & Remm (2009).

Takseertunnuste kaartide koostamiseks kasutatava *k*NN tööpõhimõte on lihtne. Eeldatakse, et näidisteks on olemas teatud hulk proovitükke või puistuid, millel on meid huvitavad tunnused mõõdetud. Neile näidistele arvutatakse asukohakoordinaatide järgi satelliidipiltidelt, aerofotodelt või aerolidari mõõtmistest tunnusvektorid, mis sisaldavad spektraalset heledust või selle teisendusi ning aerolidari punktipilve kõrgusjaotuse statistikuid. Tunnusvektoris võib olla ühe tunnusena ka geograafiline asukoht. Seejärel hakatakse kaugseireandmestikku pikselhaaval läbi vaatama ning igale pikslile otsitakse tunnusvektori järgi *k* kõige sarnasemat näidist. Sarnasuse mõõtmiseks kõige lihtsamal juhul arvutatakse näidise ja vaadeldava piksli tunnusvektorite vaheline Eukleidiline kaugus (1). Vaadeldavale pikslile omistatakse *k* kõige lähema näidise andmetest võetud soovitud takseertunnuse väärtus (tüvemaht *M*, metsa kõrgus *H*) kas aritmeetilise keskmisena või eelnevalt arvutatud kaugusega pöördvõrdeliselt kaalutuna (2, 3) (Fazakas *et al*., 1999; McRoberts, 2012). Nominaaltunnuste korral nagu näiteks puuliigi kood kasutatakse moodväärtust. Üheks olulisimaks probleemiks *k*NN meetodi puhul on saadavate hinnangute nihe (Poso *et al*., 1990; Holmgren *et al*., 2000) ehk süstemaatiline erinevus tegelikust väärtusest, mille põhjusteks peetakse kaugreiretunnuste ja maapeal mõõdetud takseerandmete vahelisi mittelineaarseid ja küllastuvaid seoseid (joonis 3) ning seda, et enamasti on kõikide tunnuste kaalud positiivsed (Fazakas *et al*., 1999). Lähima naabri klassifitseerimistehnika puhul on uuritud nii võimaliku naabrite arvu kui ka tunnusvektorite dimensionaalsuse vähendamist ning saadud ennustuste vigade hindamist. Enamik autoreid pakub naabrite arvuks $3 \leq k \leq 10$, aga näiteks McRoberts (2008, 2012) saab kogu protsessi optimeerimise järel üldiselt $k > 25$. Kaugseiretunnuste valiku osas on kasutatud geneetilisi algoritme (McRoberts, 2008; Tomppo *et al*., 2009), klasteranalüüsi (Tamm & Remm 2009), peakomponentide meetodit (Chirici *et al*., 2008) ja ekspertarvamust (Zald *et al*., 2014).

Laeva-Kursi piirkonda rajati 2013. aastal 15 × 15 km katseala (joonis 1) eesmärgiga uurida *k*NN meetodi abil mitmerindelistes

segapuistutes rinnete tüvemahu ja puistu koosseisu hindamist. Näidistena kasutati 444 proovitükki (tabel 1, joonis 2), tüvemahu hinnangute kontrollimiseks oli 89 metsa kasvukäigu proovitükki ja 2290 puistut metsakorralduse andmebaasist. Kontrollandmestikus ennustati tüvemaht mudelite abil 2013. aasta kohta, millal tehti lennukilt laserskanneerimine. Süstemaatilise allahindamise tõttu (Raudsaar *et al*., 2014) skaleeriti metsaregistri takseerkirjelduste tüvemaht (M_{IDB}) proovitükkidel mõõdetud tüvemahu (M_{tr}) vahemikku (joonis 4). Ennustustes testiti viit tunnusvektorite komplekti (tabel 2), mis koostati originaaltunnustest või nende peakomponentidest. Testiti naabrite arvu ja leiti, et võrreldes $k = 3$ on teiste variantide puhul saadud hinnangute erinevused väikesed. Uuriti kirjeldava tunnuste kaalu a_j (1) mõju tüvemahu ennustustele ALS andmekomplektil põhineva tüvemahu ennustusele andes kaalule a_j tabelis (3) toodud väärtused. Korraldati ka puistute koosseisu ennustamise katse, milles näidisalade hulgast eemaldati kuni 70% kaseenamusega proovitükidest.

Tulemustest selgus sarnaselt paljudele varasematele uuringutele, et aerolidari andmete kaasamine parandab tüvemahu ennustamise täpsust ning vähendab ennustatava tüvemahu väärtusest sõltuvat nihet nii esimese kui ka teise rinde puhul (joonised 5, 6, 7). Samas ei õnnestunud ühegi kaugseiretunnuste komplekti puhul süstemaatilist nihet tüvemahu hinnangutest kõrvaldada. Ilmnes ka, et tunnuste ruumi mõõtmete vähendamine peakomponentide meetodi abil on tõhus võte, sest näiteks kõikide sisendtunnuste (tabel 2) viis esimest peakomponenti sisaldasid 99,7% kogu informatsioonist ning ennustatud tüvemahud või puistute koosseis oli sama täpne või isegi täpsem originaaltunnuste kasutamisega võrreldes. Üsna uudseid tulemusi saadi puistute liigilise koosseisu ennustuse analüüsimisel seoses puistute vanusega. Selgus, et ennustatud koosseisus esineb süstemaatiline viga puuliigiti sõltuvalt puistute vanusest (joonis 8) ning kaskede osakaalu hinnatakse nooremates puistutes alla ja vanemates süstemaatiliselt üle, aga haabade osakaalu puhul on täpselt vastupidi. Kuna näidisalade ja neid sisaldavate metsaeraldiste andmete võrdlemisel sarnast seaduspära ei ilmnenud (joonis 9), siis tuleb sellise süstemaatilise vea põhjuseks pidada *k*NN tehnika omapära, mis on seotud tunnusvektorite info kattumisega segapuistutes. Kui näidisalade hulgast eemaldati vanusklassiti kuni 70% kaseenamusega proovitükkidest, et muuta haava ja kaseenamusega näidiste jagunemist vanuse järgi võrdsemaks, muutusid ka puuliikide ennustatud osakaalud, kuid nihked hakkasid tekkima ka eelnevalt pigem täpselt ennustatud kuuskede osakaalus (joonis 10). Selline vea ümberpaiknemine sõltuvalt näidisproovitükkide jaotusest koosseisu või vanuse järgi võib osutuda oluliseks neis uuringutes, kus kasutatakse tagasipanekuga bootstrap valimit (McRoberts, 2012) *k*NN ennustuste usalduspiiride konstrueerimiseks aladel, kus ühe pildi ulatuses esineb piirkonniti erinevaid metsi.

Received October 27, 2014, revised November 28, 2014, accepted December 10, 2014