Construction of tree species composition map of Estonia using multispectral satellite images, soil map and a random forest algorithm

Mait Lang[1,2*], Mihkel Kaha[1], Diana Laarmann[2] and Allan Sims[3]

Lang, M., Kaha, M., Laarmann, D., Sims, A. 2018. Construction of tree species composition map of Estonia using multispectral satellite images, soil map and a random forest algorithm. – Forestry Studies | Metsanduslikud Uurimused 68, 5–24. ISSN 1406-9954. Journal homepage: http://mi.emu.ee/forestry.studies

Abstract. Landsat-8 OLI and Sentinel-2 MSI images from years 2015 and 2016, a 1:10,000 digital soil map and a large number of reference samples were used with a random forest machine learning implementation in GRASS GIS to construct a tree species map for the entire territory of Estonia (42,755 km^2). Class probabilities for seven main tree species, an extra class for other species and probability of the forest cover not conforming to the forest definition were assigned for each pixel. Validation of dominant species distribution by area showed very strong correlation at county level both in state forests (R^2 = 0.98) and in private forests (R^2 = 0.93). Validation of tree species composition using harvester measurement data from 2,045 regeneration felling areas showed also very strong correlation (R^2 = 0.75) with the measured values of the proportion of coniferous trees. There was some tendency to underestimate the proportion of more common species and overestimation was found for the species with smaller proportion in the mixture. The accuracy for the proportion of deciduous species that were present in a smaller number of reference observations was substantially smaller. Validation of the results by using data from 659 large sample plots from the database of the Estonian Network of Forest Research Plots and 3,002 small sample plots from the National Forest Inventory (NFI) data base confirmed the findings based on harvester data. The NFI data revealed also a decrease of estimation error with the increase of forest age. Cohen's kappa index of agreement for main species for NFI sample plots with main species proportion equal to or greater than 75% decreased from 0.69 to 0.66 when observations with forests younger than 20 years were included in the comparison. Overall, the constructed map provides valuable data about tree species composition for the forests where no up to date inventory data are available or for the projects that require continuous cover of tree species data of known quality over the entire Estonia.

Key words: forest inventory, random forest, tree species, raster map.

Authors' addresses: [1]Tartu Observatory, Faculty of Science and Technology, University of Tartu, 61602 Tõravere, Tartumaa, Estonia; [2]Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Kreutzwaldi 5, 51014 Tartu, Estonia; [3]Forest Department, Estonian Environment Agency, Mustamäe tee 33, 10616 Tallinn, Estonia; *e-mail: mait.lang@to.ee

DOI: 10.2478/fsmu-2018-0001

Introduction

Sustainable forest management planning and forest policy making requires forest inventory data at different spatial scales and generalization levels. Individual forest stand data are used for forest management planning (Metsakorralduse, 2017). Forest policy development and general monitoring of trends is usually based on National Forest Inventories (NFI) carried out on a sampling grid of permanent sample plots (Tomppo *et al.*, 2010). Both of the forest inventory methods use remote sensing data. Stands are delineated for the forest management planning using aerial photographs (Spurr, 1948). Maps of wood volume estimates that are constructed with nonparametric estimation methods e.g. k-nearest neighbour (*k*NN) using sample plot measurements and feature variables from multispectral satellite images are some of the outputs of NFIs (McRoberts & Tomppo, 2007). In Estonia Tamm & Remm (2009) used reference set observations taken from a forest management inventory database (FIDB), Landsat ETM+ images and a 1:10,000 digital soil map data for machine learning-based construction of standing wood volume maps and obtained root mean square error (RMSE) 87.0 m^3 ha^{-1} (42%) at pixel level and 74.6 m^3 ha^{-1} (36%) at stand level.

Wood volume data are important for the estimation of carbon storage and estimation of timber, but tree species composition data are required for biodiversity assessment (Laarmann *et al.*, 2009; McRoberts *et al.*, 2012), satellite-based estimation of net primary production (Zhao *et al.*, 2011; Lang *et al.*, 2017), ecosystem models (Duveneck *et al.*, 2015), and also for purposes of monitoring and forest industry planning. Tree species composition of a forest stand is a vector of the relative proportions of individual species stemwood volume from total stemwood volume of the stand. However, maps with up-to-date inventory estimates in the Forest Inventory Data Bases (FIDB) cover usually only the forests where the owner is interested in management and the number of NFI sample plots per unit area is only sufficient for estimating regional averages. For spatially continuous estimates, a machine learning approach with spatial feature variables (multispectral images, airborne lidar data, soil maps) can be used to construct tree species composition maps at medium spatial resolution (20–30 m) for all forests in the region. Decision trees-based random forest-type (RF) methods have been successfully used for tree species classification and land cover mapping (Yang *et al.*, 2014; Barrett *et al.*, 2016).

In this study we used a RF implementation (r.learn.ml) in GRASS GIS (GRASS Development Team 2017), Landsat-8 Operational Land Imager (OLI) and Sentinel-2 Multispectral Instrument (MSI) images, a 1:10,000 digital soil map and a large number of reference samples drawn from an FIDB to construct a tree species map for all of Estonia. The map was then validated using county level statistics, harvester measurements from regeneration felling stands, samples from the database of Estonian Network of Forest Research Plots (Kiviste *et al.*, 2015) and also a set of NFI sample plots.

Material and Methods

The location

The study area (Figure 1) included the entire terrestrial territory of Estonia (42,755 km^2) except for Ruhnu, a small and distant island in the Baltic Sea. About half (53.2%) of the Estonian terrestrial territory is forest land, which is 51% publicly owned by state or municipalities while the remainder belongs to private forest owners; a small part (46,341 ha) was retained by the state after land restitution following the collapse of the Soviet Union (Raudsaar *et al.*, 2017, Valgepea & Maamets, 2017). The most widespread forest trees in Estonia are European aspen (*Populus tremula* L.), silver birch (*Betula pendula* Roth), downy birch

Figure 1. The study area covers the entire territory of Estonia. The squares mark eight image subsampling regions that overlap partially. County borders are drawn on the map.

Joonis 1. Pilditöötlus tehti väiksemate ristkülikukujuliste serviti kattuvate piirkondade (katkendjooned) kaupa. Kaardile on kantud ka maakondade piirid.

(*B. pubescens* Ehrh.), Norway spruce (*Picea abies* (L.) Karst.), black alder (*Alnus glutinosa* (L.) Gaertn.), grey alder (*A. incana* (L.) Moench), Scots pine (*Pinus sylvestris* L.) and common ash (*Fraxinus excelsior* L.). These tree species grow in different mixtures in Estonia. On fertile soils Norway spruce is also common in the mid-story and the understory of the forests. Due to the historical background state owned forests stands are dominated by Scots pine, birch spp. and Norway spruce and in private land the forest stands are dominated by birch spp. followed by Scots pine and grey alder (Raudsaar *et al.*, 2017). This difference is caused by the tendency for natural regeneration of fast growing broadleaf deciduous species after regeneration fellings (Raudsaar *et al.*, 2014) and the large share of abandoned agricultural private land where fast growing broadleaf deciduous trees do occur in the first order.

Ancillary data and feature variables

Satellite imagery from Landsat-8 Operational Land Imager (OLI) and Sentinel-2 Multispectral Instrument (MSI) (Table 1) were downloaded from the USGS GloVis portal (https://glovis.usgs.gov) and from the Copernicus Open Access Hub (https://scihub.copernicus.eu). We used top-of-atmosphere radiances and did not apply atmospheric correction. Using cloud-free sub-regions of the images it was possible to cover the entire country and pairs of images on different dates provided information on phenology as this has been found informative for mapping species composition (Wilson *et al.*, 2012).

All the images were transformed into the Estonian base map coordinate system (EPSG:3301) using GDAL tools (www.gdal.org). A pixel size of 25 m was used as a compromise between the original spatial resolution of the images and the large volume of data. Nearest neighbour resampling was used for the 30 m resolution Landsat-8 OLI (USGS, 2016) bands and for the Sentinel-2 MSI bands that have 20 m resolution. For the Sentinel-2 MSI bands with 10 m resolution (SUHET, 2015), the averaging of source pixel values was used for raster image resampling. Cloud and cloud shadow areas were delineated manually. Both of the scanners have a special channel near to the water absorption band in the electromagnetic spectrum that was useful for haze and cirrus detection. The entire territory of Estonia was subdivided into partially overlapping regions (Figure 1) according to forest growth conditions and to establish data processing units with sufficient counts of reference samples. After cutting the individual Landsat and Sentinel images according to the regional borders for data processing there were 85 image combinations with sufficiently large area and each containing a single image or a pair of images with phenology effect. In the central region the number of image combinations was the greatest (15), but in the western Estonia and in the islands only 7 image combinations could be constructed.

Table 1. The list of satellite images used in this study; Sentinel-2 images are referenced by orbit number and Landsat-8 images using World Reference System-2 path and row.

Tabel 1. Satelliidipiltide nimekiri. Asukoht on Sentinel-2 puhul orbiidi number ja Landsat-8 puhul World Reference System-2 orbiit (path) ja pildirida (row).

Satellite and scanner *Satelliit ja skanner*	Location *Asukoht*	Date *Kuupäev*
Landsat-8 OLI	185;019	08.08.2015
Landsat-8 OLI	185;019	24.08.2015
Landsat-8 OLI	185;019	06.05.2016
Landsat-8 OLI	186;019	12.06.2015
Landsat-8 OLI	186;019	13.05.2016
Landsat-8 OLI	186;019	14.06.2016
Landsat-8 OLI	187;019	21.07.2015
Landsat-8 OLI	188;019	25.05.2015
Landsat-8 OLI	188;019	11.05.2016
Landsat-8 OLI	189;019	03.07.2015
Sentinel-2 MSI	R036	04.08.2015
Sentinel-2 MSI	R036	14.08.2015
Sentinel-2 MSI	R036	24.08.2015
Sentinel-2 MSI	R036	10.05.2016
Sentinel-2 MSI	R036	28.08.2016
Sentinel-2 MSI	R079	07.08.2015
Sentinel-2 MSI	R079	17.08.2015
Sentinel-2 MSI	R079	13.05.2016
Sentinel-2 MSI	R122	06.05.2016
Sentinel-2 MSI	R122	24.08.2016
Sentinel-2 MSI	R122	13.09.2016
Sentinel-2 MSI	R136	21.08.2015
Sentinel-2 MSI	R136	30.09.2015
Sentinel-2 MSI	R136	27.04.2016
Sentinel-2 MSI	R136	07.05.2016

Processing data by geographical sub-regions has been used to decrease the possibility of erroneously predicting a species outside its realistic region of occurrence (Duveneck *et al.*, 2015). We used a similar approach (Figure 1) but due to the variability of the Estonian forested landscape we included also data from the 1:10,100 national

soil map (Mullakaardi, 2001). The soils database was downloaded from the website (www.maaamet.ee [01.02.2017]) of the Estonian Land Board. Each soil polygon is associated with a data record containing the soil type and morphological information. For the machine learning procedure the soils were grouped according to Table A1.1 using the pedo-ecological schema of normally developed mineral soils (see Figure 2 in Kõlli *et al*., 2004). The vector map of soils was rasterized, each pixel was assigned the soil group code, and the soil group code was used as a categorical variable in the RF estimation procedure.

Reference set data

While NFI data are used as reference sets in many studies for machine learning we used sample plots from the Estonian NFI only for validating our results. The NFI sampling grid is designed so that each annual sampling unit corresponds to 1,000 ha in land category-based estimates. Starting from 2014, the grid was modified and the estimated number of yearly measured sample plots is now 5,600. Since the share of forest land is 53%, there will be about 200 sample plots per county with possible tree species composition data. This is not much when considering that in addition to species composition the spectral signatures of forests also depend on stand age, leaf area index, stand density, and ground vegetation (Nilson & Peterson, 1994). By incorporating sample plots from a larger area comprised of regions with different growth conditions and from several sampling years, this increases the number of observations but also increases the risk of mixing samples with similar spectral signature but different species composition. At the same time, reference samples for model fitting and validation of the results are lost in places where clouds, cloud shadows and haze influence pixel values in the satellite images. Since NFI sample plots are small (radius 7 to 10 m), their positioning errors combined with raster image re-sampling errors introduce substantial random errors in the spectral signatures of the sample plots. Finally, the locations of NFI sample plots do not follow forest stands, but are determined by the sampling grid. Hence, sample plots can be located near to stand borders and thereby have a mixed spectral signature.

We used for our machine learning procedure the forest management inventory database from the Estonian Environmental Agency (Forest database, 2016). The database contains a 1:10,000 map of stand borders and mensurational data (forest age, stand height, basal area, stand relative density, site class, site type, wood volume, *etc*.) used for forest management planning. For each stand element (a tree species growing in particular social layer) its proportion is given according to wood volume. Although total wood volume is known to be underestimated in the database (Lang *et al*., 2014; Arumäe & Lang, 2016), the distribution of wood among species and thereby the tree species composition is usually reliable. The only exception is a small systematic underestimation of Norway spruce in state owned forests, according to volume measurements made at time of harvest (Tavo Uuetalu, The Estonian State Forest Management Company, personal communication).

A copy of the FIDB was obtained from the Estonian Environmental Agency on 13 February 2017. The FIDB data were preprocessed similar to Lang *et al*. (2016) to extract sufficiently large and compact forest stands for a reference set and to exclude outdated FIDB records and the polygons with substantial variability in pixel values. As the RF estimation procedure in GRASS is pixel-based, the within stand variability described by variograms is technically complicated to use. However, training data can be prepared separately from the estimation procedure and we used mean values of pixels located near to stand polygon centroids instead of single, nearest to centroid pixel value.

Firstly, a subset of forest land parcels inventoried since 2014 with area between 1 and 10 hectares was extracted. The extracted polygons were buffered 30 m towards the inside and the areas under large ditches were cut out. Irregular polygons were then deleted from the selection. For each remaining polygon the mean radiance value was calculated for each band in the satellite images and only the polygons for which at least 16 pixels were extracted were kept.

Secondly, the selection was filtered using their spectral radiance. Parcels were retained if the ratio of standard deviation to mean in the near infrared radiance (NIR) bands (OLI5, MSI08) was less than the 97.5th percentile of the population value. This filter excluded internally variable polygons. Next, the remaining observations with possible disturbances from 2015 to 2016 were identified according to radiance changes in the blue (OLI2, MSI02) and NIR bands (OLI5, MSI08). For this the reference observations were grouped by main species and those deviating more than four residual standard errors from a linear regression model between radiances from different dates were excluded as disturbed.

Thirdly, the concordance of spectral radiance on forest age and wood volume of remaining observations were analysed. Since the images were taken over two years and the forest inventory records were also from later dates than the last image, some polygons had small radiance in the short-wave infrared (SWIR) bands (OLI6, MSI11) characteristic of old stands but the forest age was zero. This hints at outdated or conflicting data, since young stands are brighter then old stands (Nilson & Peterson, 1994). All the observations were removed that had zero age and less than average radiance of the 1 to 6 year-old forests observations in the SWIR bands. Finally, all the stands older than 20 years or with wood volume over 50 m^3 ha^{-1} were grouped according to main species and outliers were identified by their spectral radiance in 10 year age classes and 50 m^3 ha^{-1} volume classes. The stands with radiance in red, NIR or SWIR bands deviating more than three standard deviations from the class mean were excluded. The procedure was repeated three times. Some outliers in the classes with a small number of observations (very old stands) were identified visually from scatter plots of wood volume and spectral radiance values. About 480 outliers were later detected and excluded when feature variable values from a 3 × 3 pixel window around polygon centroids were calculated for the random forest algorithm. The count of reference samples after all the outliers were removed was 102,291.

The random forest model fitting and map construction

Random forests is a machine learning algorithm for classification that corrects for overfitting of the training set (Breiman, 2001). The random forest (RF) classification algorithm (r.learn.ml) in GRASS has the following hyper-parameters: number of feature variables during node splitting N_{feat}, maximum tree depth H_{max}, minimum number of samples required for node splitting N_{split}, minimum number of samples for leaf node N_{leaf}, and number of estimators N_{trees}. The values for the hyper-parameters are recommended to fit according to the user guide. Since the model construction involves random sampling of features during building of the trees, the results are dependent on the random number generator initial state I_{rand}. During some initial tests we found that the algorithm was influenced by the distribution of observations between classes similar to *k*NN as found by Lang *et al.* (2014, 2016). We found that our estimates were more reliable with automatic balancing switched on. We also used the standardisation option of feature variables. The number of permutations of feature variable values during model construction was set to 10 (default value). The number of estimators was fixed to default value $N_{\text{trees}} = 100$.

The machine learning module includes a cross-validation component. We used

non-spatial nested cross-validation with the reference set split to 2 folds for smaller sets of observations and up to 5 folds for larger reference subsets determined by the useful area of image pairs. In this study the cross-validation of a single image or image pair based results was not of interest, however, cross-validation was important for optimisation of the algorithm parameters and selection of informative feature variables.

The RF classification algorithm predicts a class code and also probabilities of all classes for each pixel in the target set. These probabilities can be used as reliability estimates (Barrett *et al*., 2016). However, considering that spectral signature of a forest stand is a linear mixture of the reflectance of the trees, these probabilities can also be interpreted as species composition estimates for forest stands assuming that different species have different optical properties. In this study we used seven classes corresponding to the most widespread tree species in Estonia and one class for other tree species. A separate class was used for the reference observations located in recent regeneration felling areas where the tree canopy did correspond to the forest definition in Estonia.

The estimation procedure was performed in four steps for each image or image pair:

1. search for informative feature variables;
2. preparation of reference sample data;
3. search for the model hyper-parameter optimum values; and
4. imputation of target set pixel values.

In the first step the hyper-parameters of the RF algorithm were set to $N_{\text{feat}} = 0$ (automatic), $H_{\max} = 27$, $N_{\text{split}} = 30$, $N_{\text{leaf}} = 8$ and four estimates were imputed using $I_{rand} \in \{3,9,41,87\}$. Feature variable values were extracted from centroid pixels of the reference polygons. The feature importance's for each run were obtained from nested cross-validation and averaged. Soil data were almost always ranked at as one of most important feature variables and followed usually by the NIR bands. The six to ten most informative spectral bands were selected for the RF model training and estimation. In the second step a training data table was created by sampling pixels located closer than 36 m to the reference stand centroid position. Spectral radiance was averaged and mode value of the pixels in soil map codes was used.

In the third step the RF model was fitted to the training data to find optimal values for the hyper-parameters. We fixed the maximum number of features for node splitting to two features less than the number of features found during the optimization step (not all features are always required), N_{leaf} was fixed to a value between 5–10 depending on the sample size, N_{split} was usually set to $3N_{\text{leaf}}$, random state was $I_{\text{rand}} = 1$ and maximum tree depth value was the free parameter searched from the range of 15–50.

Finally, imputation of the target set pixel values was carried out for $I_{rand} \in \{1,3,6\}$. The procedure yielded 255 estimates considering all the sub-regions shown in the Figure (1); i.e., three I_{rand} states for each of the 85 image combinations.

For each reference set pixel the probabilities of classes were averaged from available estimates; this produced 9 raster layers that covered all of Estonia. The map layers correspond to the probabilities of tree species in the composition and one layer contained the probability that the pixel does not correspond to forest stand definition (tree layer was too sparse or trees were too small). The species composition estimates for each pixel were calculated by scaling the tree class probabilities to sum to 100 excluding the non-forest probability class. Finally, the class code with the highest probability was then indicated in a separate layer of the tree species composition map. Using the Estonian 1:10,000 base map, the pixels with highly improbable occurrence of tree cover were assigned a no data flag. The no data flag was also assigned to agricultural land where the records of the Estonian Agricultural Registers and Information

Board indicated the landowner applied for a subsidy in 2009–2011.

Validation of the species composition map

The validation of the tree species composition map was carried out using four data sources. The first validation dataset was the area distribution of inventoried stands by dominant tree species in counties according to official national statistics (Raudsaar *et al.*, 2017). The pixel distribution of the dominant species layer was calculated from the species composition map. The second validation dataset contained timber volume measurements made by harvesters in regeneration fellings that was made available by the Estonian State Forest Management Company. From this dataset we selected 2,045 records corresponding to the stands that contained at least 16 pixels and where more than 85% of the pixels were assigned a tree species class as the most probable and less than 20% of the harvested timber (fuelwood, *etc.*) was not assigned to a tree species. The mean proportion of each tree species in the species composition map was calculated for each stand and stand level data were compared with the harvester measurements.

The third validation dataset was extracted from the database of Estonian Network of Forest Research Plots (ENFRP) (Kiviste *et al.*, 2015) from the list of sample plots measured from 2012 to 2015. The sample plots have a radius ranging from 15 m to 30 m depending on forest age, thus a sample plot could cover an area larger than the 25 m pixel. Mean age of the forests was 69 years ranging from 17 to 243 years. The sample plots were established in homogeneous parts of forest stands and therefore are representative also for their near vicinity. After checking for possible stand replacing disturbances using orthophotos and according to the forest age and brightness relationship, the number of suitable observations was 659 observations.

A subset from the Estonian National Forest Inventory (Adermann, 2010) database was used for a fourth set of validation tests of the stand map. These sample plots have a radius of 7 m or 10 m and the error in coordinate values is usually less than 45 m. A subset of 3,002 sample plots was extracted to analyse species composition. The selection criteria were as follows: the plot was not near to roads or ditches, the wood volume was greater than 5 m^3 ha^{-1} and the probability of the non-forest (NFD) class was less than 10%. Each plot was related to the imputed values of the nearest pixel of the sample plot location. For each sample plot Euclidean distance between the measured and predicted species composition vectors was calculated as

$$D_{spc} = \sqrt{\sum_{i=1}^{n}(k_{i,SMI} - k_{i,map})^2}, \quad (1)$$

where $k_{i,SMI}$ is the proportion of the species i in the NFI data and $k_{i,map}$ is the predicted proportion of the i^{th} species. Dependence of on the forest age was analysed.

For estimates of categorical variables as land cover types it is common to report Cohen's kappa index of agreement. Tree species proportion, however, is a complex ratio variable. It is reasonable to calculate Cohen's kappa only for the validation sample plots where proportion of dominating species is sufficiently large to classify the observations as pure stands of particular tree species. From the NFI data we first selected the sample plots where dominating species proportion was equal or greater than 75%. A second validation subsample of pure stands was created by excluding the stands with age less than 20 years. All statistical analyses were carried out in R (R Core Team, 2016).

Results and Discussion

The total area of the map pixels with tree species composition is 2.26 million hectares that is about 8% more than the official estimate (Raudsaar *et al*., 2017) of forested forest land area of 2.09 million hectares. The difference is related to distinction of forest land from bush and bog land categories, and interpretation of land cover for small wooded land patches. The comparison of the main species distribution in inventoried forest stands at the county level indicated a high correlation between the predicted values and national statistics (Figure 2). For state forests (862,136 ha) the determination coefficient R^2 was 0.98 and for private forests (803,525 ha) R^2 was 0.93. The random forest algorithm-based estimates showed a larger share of grey alder and birch stands in the private forests and a greater share of Scots pine stands in state forests, similar to national inventory statistics. This result indicates that the constructed map (Figure A3.1) of tree species composition (available from Tartu Observatory web page (www.to.ee) and upon request from the corresponding author) can be used for the rest of the forest land (410,778 ha) for which there are no records in the FIDB.

The aggregated estimates of main species at the county level indicated that there are no substantial shortcomings in the data processing. However, our objective was to obtain a map of tree species composition, not only the main species. The comparison of predicted proportion of tree species in stands with the harvester measurements showed the strongest correlation for Scots pine followed by Norway spruce, birch, and European aspen (Table 2). The predicted proportion of coniferous trees had very strong correlation (R^2 = 0.75) with the measured value. However, there was also substantial scatter in the relationship (Figure 3) and the gain of the expected linear model was only 0.67. There was a characteristic lack-of-fit with large values

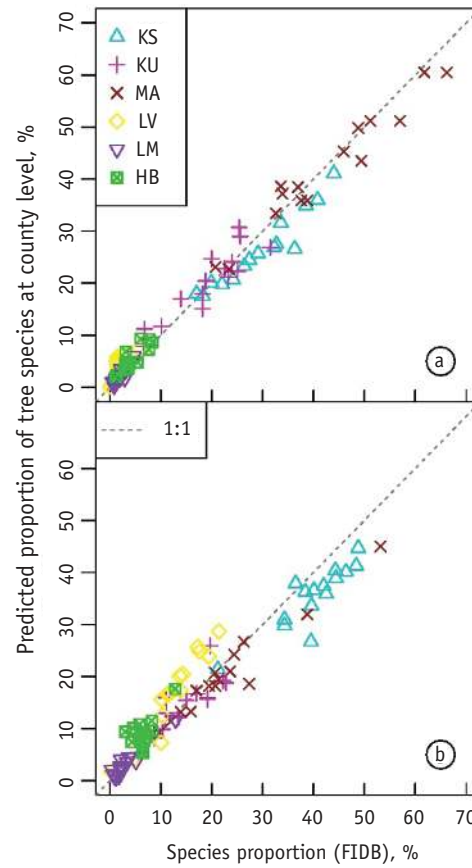

Figure 2. County level comparison of the predicted proportion of the main species to inventoried forest stands (*Raudsaar et al. 2017*) in state forests (a) and private forests (b). The abbreviations in the legend are: HB (*Populus tremula* L.), KS (*Betula pendula* Roth, *B. pubescens* Ehrh.), KU (*Picea abies* (L.) Karst.), LM (*Alnus glutinosa* (L.) Gaertn.), LV (*A. incana* (L.) Moench), MA (*Pinus sylvestris* L.), SA (*Fraxinus excelsior* L.).

Joonis 2. Korraldatud puistute jagunemine enamuspuuliigi järgi maakonniti (a) riigimetsas ning (b) erametsades puistuplaani (y-telg) ja aastaraamatu Mets 2016 (Raudsaar et al., 2017) järgi.

underestimated and small values overestimated. It appeared (Table 2) also that the mean proportion of the Norway spruce, Scots pine and birch was underestimated and the proportion of less common species

Table 2. Mean proportion of species in 2,045 forest stands based on harvester measurements and the constructed map. S_e is standard error. Harvester measurements are used for the independent variable in the linear model y = a + bx.

Tabel 2. Puuliikide osakaalude hinnangud 2045 puistus harvestermõõtmise ja puistuplaani järgi. S_e on standardviga. Seose lineaarmudelis (y = a + bx) on argumendiks harvestermõõtmistel saadud puuliigi osakaal.

Species *Puuliik*	Harvester / *Harvester*		Stand map / *Puistuplaan*		Linear model / *Lineaarseose parameetrid*		
	Mean	S_e	Mean	S_e	a	b	R^2
Silver and downy birch	19.8	0.4	15.8	0.2	9.5	0.32	0.55
Norway spruce	37.7	0.5	24.7	0.3	12.0	0.34	0.44
Scots pine	28.7	0.7	27.7	0.4	13.1	0.51	0.72
Grey alder	0.0	0.0	5.6	0.1	5.6	2.72	0.01
Black alder	0.3	0.0	7.1	0.1	6.7	1.31	0.12
European aspen	8.9	0.3	10.3	0.1	8.2	0.24	0.42
Common ash	0.0	0.0	3.9	0.1	3.9	4.13	0.00
Other tree species	4.6	0.1	4.8	0.1	3.8	0.21	0.20

tended to be overestimated. Estimation errors can cause biased values because the proportion of species cannot be negative or greater than 100%. Additionally, the random forest algorithm only predicted the probability of classes (main species) and the scope of this project did not include reprogramming of the RF implementation to access directly the data vectors of the reference observations. Also, the target set pixel values were calculated as mean values of several estimates based on available image combinations and three selected random state values.

The mean proportion of Norway spruce was substantially greater in the harvester dataset than in the constructed map, similar to the apparent underestimate in the forest inventory database (FIDB) compared to the harvester data. Since the FIDB was used to draw reference set observations for the random forest algorithm the imputed target set pixel values in the constructed map were probably influenced by the possible bias in the FIDB. However, it was not possible to identify the true causes of the observed systematic difference that exists between the harvester measurements and the constructed map in the proportion of Norway spruce and birch in tree species composition in forest stands.

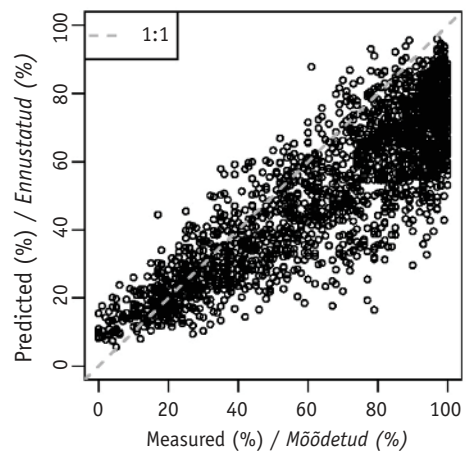

Figure 3. Measured (harvesters) and predicted proportion of coniferous trees in 2,045 forest stands having undergone regeneration felling.

Joonis 3. Uuendusraietel (2045 puistus) harvestermõõtmiste põhjal saadud okaspuude osakaal puistus võrrelduna puistuplaanil oleva hinnanguga.

While the harvester measurement data represented only old forest stands, the sample plots from the ENFRP dataset (Kiviste *et al*., 2015) included also younger stands. There was a strong correlation (R^2 = 0.79) between the measured and predicted proportion of evergreen coniferous trees, similar to harvester dataset, but the gain of the expected linear relationship was greater (0.76). Nevertheless, there was also rather large scatter at the sample plot level (Figure 4). The validation dataset confirms that the proportions of Norway spruce, Scots pine and birch were underestimated and the proportion of less common species was overestimated in the target set pixels (Table 3).

The validation based on NFI sample plots (Table 4) showed results similar to the other validation data sets. However, the coefficient of determination R^2 between the measured and predicted proportion of coniferous trees at the sample plot level was only 0.64. A similar decrease in R^2 also was present at the species level compared to the other validation data sets. The main reason for the decreased correlation likely was the smaller plot size with respect to pixel size and errors in the spatial location of the sample plots. However, the NFI dataset covered the entire age range of forests and this enabled a study of possible dependence of the estimation errors in relation to forest age. There was a decrease (slope –0.21 in linear model, p-value < 0.001) (Figure 5) in the estimation error depending on forest age (R^2 = 0.1, p-value < 0.001 at 3,000 degrees of freedom). There was no age dependence in the number of tree species in the NFI sample plots as indicated by non-significant slope (p-value > 0.2) and R^2 = 0 of the relationship.

Table 3. Mean proportion of species in 659 sample plots from ENFRP (Kiviste *et al*., 2015) database and the constructed map. S_e is standard error. Sample plot measurements are used for the independent variable in the linear model y = a + bx.

Tabel 3. Puuliikide osakaalude hinnangud 659 kasvukäiguproovitükil (KKPRT) (Kiviste et al., 2015) mõõtmiste ja puistuplaani järgi. S_e on standardviga. Seose lineaarmudelis (y = a + bx) on argumendiks proovitükkidelt saadud puuliigi osakaal.

Species *Puuliik*	ENFRP sample plots / *KKPRT proovitükid*		Stand map / *Puistuplaan*		Linear model / *Lineaarseose parameetrid*		
	Mean	S_e	Mean	S_e	a	b	R^2
Silver and downy birch	16.3	0.95	10.6	0.39	5.38	0.32	0.61
Norway spruce	32.7	1.28	27.9	0.85	9.21	0.57	0.73
Scots pine	44.2	1.59	40.5	1.08	13.55	0.61	0.81
Grey alder	0.6	0.11	4.0	0.17	3.89	0.25	0.03
Black alder	1.3	0.22	4.6	0.19	4.07	0.43	0.24
European aspen	3.5	0.43	6.1	0.22	5.03	0.3	0.34
Common ash	0.1	0.04	2.9	0.09	2.81	0.44	0.03
Other tree species	1.3	0.2	3.4	0.11	3.19	0.17	0.1

Table 4. Mean proportion of species in 3,002 NFI sample plots and the constructed map. S_e is standard error. Sample plot measurements are used for the independent variable in the linear model y = a + bx.

Tabel 4. Puuliikide osakaalud 3002 statistilise metsainventuuri (SMI) proovitükil välimõõtmiste ja puistuplaani järgi. S_e on standardviga. Seose lineaarmudelis (y = a + bx) on argumendiks proovitükkidelt saadud puuliigi osakaal.

Species *Puuliik*	NFI sample plots / *SMI proovitükid*		Stand map / *Puistuplaan*		Linear model / *Lineaarseose parameetrid*		
	Mean	S_e	Mean	S_e	a	b	R^2
Silver and downy birch	26.0	0.57	17.6	0.21	11.8	0.22	0.35
Norway spruce	23.0	0.53	18.4	0.31	9.98	0.37	0.4
Scots pine	29.3	0.7	26	0.47	10.27	0.54	0.65
Grey alder	6.5	0.36	8.7	0.17	6.87	0.29	0.35
Black alder	5.4	0.31	9.0	0.13	7.87	0.2	0.23
European aspen	5.8	0.32	10.1	0.13	9.22	0.15	0.15
Other tree species	4.1	0.25	10.3	0.13	9.58	0.17	0.11

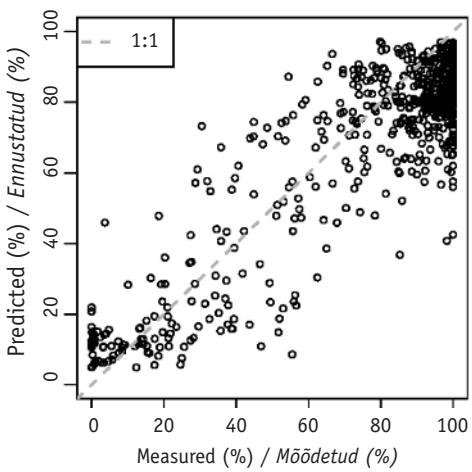

Figure 4. Measured and predicted proportion of coniferous trees in sample plots taken from ENFRP database (Kiviste *et al., 2015*).

Joonis 4. Okaspuude osakaal puistu koosseisus metsa kasvukäigu proovitükkide andmete ja puistuplaani järgi.

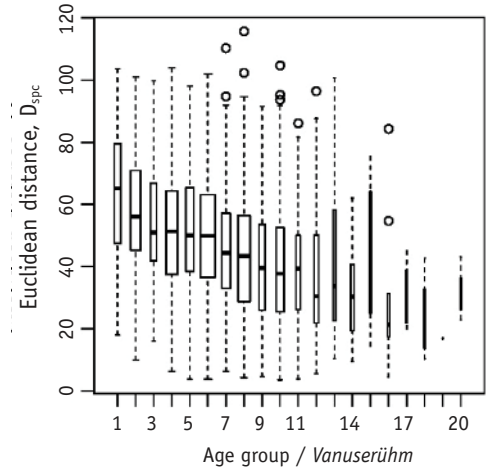

Figure 5. Mean Euclidean distance (1) between the vectors of measured and predicted species composition in the NFI sample plots. The age groups correspond to 10 years interval.

Joonis 5. Takseeritud ja ennustatud puistukoosseisude eukleidiline kaugus (1) statistilise metsainventuuri proovitükkidel 10-aastaste vanusrühmade kaupa.

Overall accuracy of the main species estimates was 75.5% for the 1,529 NFI sample plots where a single tree species proportion was equal or greater than 75%. Mean value of Cohen's kappa index of agreement was 0.66. Scots pine stands were the most accurately discriminated followed by Norway spruce stands. Separation of deciduous species was less accurate (Table A2.1). This is similar to the results obtained from the analysis of species composition. By excluding stands less than 20 years-old the overall accuracy increased to 78.4% (Table A2.2) and kappa increased to 0.69 for the subsample of 1,354 NFI sample plots. The increase in accuracy is small but consistent with the previous analysis (Figure 5) that showed increased estimation accuracy of tree species composition in older stands. We used class membership probabilities of RF machine learning procedure as estimates of tree species proportions in the species composition. The underlying assumption was that the spectral signature of a forest depends linearly on tree species composition. The validation results indicated that the constructed map of tree species composition provided reliable estimates of the main tree species in all counties in Estonia. Discrimination of deciduous tree species proportions in a mixture was less accurate than Norway spruce and Scots pine proportion, which is related to the differences in their spectral signatures. Main species estimation accuracy at the pixel level for NFI sample plots with dominant species proportion of 75% and more was 75.5% and increased when young stands were excluded. However, predictions of species proportions in composition at the stand or pixel level have a lack-of-fit characterized by an underestimation of larger values and overestimation of smaller values. While our assumption was justified that the class probabilities predicted by the random forest procedure may be used as linear proportions of species, as shown by correlation analyses, there are options to improve precision. For example, there was always about 2 to 3% probability for each class in the imputations. Considering that there were 9 classes in the dataset, the results have always about 18 to 27% noise. This noise could be reduced by direct processing of the data vectors of the reference samples from leaf nodes of the decision trees in the random forest model. However, this requires modifications in the software of the GRASS machine learning module. The forest age dependence of the predicted species composition indicates that a two-stage approach could be tested in future studies; inclusion of canopy height information from airborne laser scanning (ALS) also may be useful to separate forests by age. It is also possible that calibration of the class membership probabilities can improve the accuracy (Niculescu-Mizil & Caruana, 2005), however, the procedure requires an additional independent set of observations.

Conclusions

In this study we processed freely distributed multispectral satellite images using a Random Forest implementation in free software GRASS and constructed the first high spatial resolution map of tree species composition for Estonia. Validation of the map showed good discrimination between deciduous broadleaf and evergreen coniferous species, but separation and estimation of proportion of deciduous broadleaf species was less accurate and this has to be targeted in further studies. Overall, the constructed map provides valuable data for the forests where no up-to-date inventory data are available or for projects that require continuous cover of tree species data of known quality over all of Estonia.

Acknowledgements. The Estonian State Forest Management Centre provided financial support for the study. The Estonian Network of Forest Research Plots is supported by the Estonian State Forest Management Centre and the Estonian Environ-

mental Investment Centre. The Estonian Land Board released digital soil map for public use. Landsat-8 OLI and Sentinel-2 MSI images were made available by the USGS and Copernicus framework. Discussions with Dr. Kalle Remm helped to improve our understanding about machine learning. Prof. Andres Kiviste suggested statistical tests and commented on early versions of the manuscript. Dr. John Stanturf commented the manuscript and edited the English language. We thank anonymous reviewers for comments that helped us to improve the quality of the manuscript. Authors acknowledge also GRASS GIS developers for their efforts to create free software.

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. 2016. ALS-based wood volume models of forest stands and comparison with forest inventory data. – Forestry Studies / Metsanduslikud Uurimused, 64, 5–16.

Barrett, B., Raab, C., Cawkwell, F., Green, S. 2016. Upland vegetation mapping using Random Forests with optical and radar satellite data. – Remote Sensing in Ecology and Conservation, 2(4), 212–231.

Breiman, L. 2001. Random forests. – Machine Learning, 45(1), 5–32.

Duveneck, M.J, Thomson, J.R., Wilson, B.T. 2015. An imputed forest composition map for New England screened by species range boundaries. – Forest Ecology and Management, 347, 107–115.

Forest database. 2016. State register for accounting of forest resource (Metsaressursi arvestuse riikliku registri põhimäärus). Riigi Teataja, RT I, 12.01.2016, 2. (In Estonian).

GRASS Development Team. 2017. Geographic Resources Analysis Support System (GRASS) Software, Version 7.2.2. Open Source Geospatial Foundation. [WWW document]. – URL http://grass.osgeo.org. [Accessed 1 March 2018].

Kiviste, A., Hordo, M., Kangur, A., Kardakov, A., Laarmann, D., Lilleleht, A., Metslaid, S., Sims, A., Korjus, H. 2015. Monitoring and modeling of forest ecosystems: the Estonian Network of Forest Research Plots. – Forestry Studies / Metsanduslikud Uurimused, 62, 26–38.

Kõlli, R., Asi, E., Köster, T. 2004. Organic carbon pools in Estonian forest soils. – Baltic Forestry, 10(1), 19–26.

Korjus, H., Põllumäe, P., Kiviste, A., Kangur, A., Laarmann, D., Sirgmets, R., Lang, M. 2017. Online streaming public participation in forest management planning. – Forestry Studies / Metsanduslikud Uurimused, 66, 5–13.

Laarmann, D., Korjus, H., Sims, A., Stanturf, J.A., Kiviste, A., Köster, K. 2009. Analysis of forest naturalness and tree mortality patterns in Estonia. – Forest Ecology and Management, 258S, S187–S195.

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.

Lang, M., Gulbe, L., Traškovs, A., Stepčenko, A. 2016. Assessment of different estimation algorithms and remote sensing data sources for regional level wood volume mapping in hemiboreal mixed forests. – Baltic Forestry, 22(2), 283–296.

Lang, M., Kõlli, R., Nikopensius, M., Nilson, T., Neumann, M., Moreno, A. 2017. Assessment of MODIS NPP algorithm-based estimates using soil fertility and forest inventory data in mixed hemiboreal forests. – Forestry Studies / Metsanduslikud Uurimused, 66, 49–64.

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

McRoberts, R.E., Winter, S., Chirici, G., LaPoint, E. 2012. Assessing forest naturalness. – Forest Science, 58(3), 294–309.

Metsakorralduse. 2017. Forest inventory act. (Metsa korraldamise juhend). – RT I, 22.02.2017, 11. (In Estonian).

Mullakaardi. 2001. The fine-scale map of Estonian soils. (Vabariigi digitaalse suuremõõtkavalise mullastiku kaardi seletuskiri). Maa-amet, Tallinn. [WWW document]. – URL http:// geoportaal.maaamet.ee/est/Andmed-ja-kaardid/Mullastiku-kaart-p33.html [Accessed 16 April 2016]. (In Estonian).

Niculescu-Mizil, A., Caruana, R. 2005. Predicting good probabilities with supervised learning. – Proceedings of the 22nd International Conference on Machine Learning, August 7–11, 2005, Bonn, Germany, 625–632.

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

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [WWW document]. – URL https://www.R-project.org/ [Accessed 1 March 2018].

Raudsaar, M., Pärt, E., Adermann, V. 2014. Forest resources. – Yearbook Forest 2013. Keskkonnaagentuur, Tartu. p. 37.

Raudsaar, M., Sims, A., Timmusk, T., Pärt, E., Nikopensius, M. 2017. Forest resources. – Raudsaar, M., Siimon, K.-L., Valgepea, M. (eds.). Yearbook forest 2016. Keskkonnaagentuur, Tartu, 18–81.

Spurr, S.H. 1948. Aerial photographs in forestry. New York, Ronald Press.

SUHET. 2015. Sentinel-2 user handbook. ESA standard document. Issue 1, rev 2. 64 pp.

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., 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.

USGS. 2016. Landsat 8 (L8) data users' handbook. Version 2.0. Department of interior, U.S Geological survey. 98 pp.

Valgepea, M., Maamets, L. 2017. Forest ownership. – Raudsaar, M., Siimon, K-L., Valgepea, M. (eds.). Yearbook forest 2016. Keskkonnaagentuur, Tartu, 82–105.

Wilson, B.T., Lister, A.J., Riemann, R.I. 2012. A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. – Forest Ecology and Management, 271, 182–198.

Yang, X., Rochdi, N., Zhang, J., Banting, J., Rolfson, D., King, C., Staenz, K., Patterson, S., Purdy, B. 2014. Mapping tree species in a boreal forest area using RapidEye and LiDAR data. 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, 2014, 69–71. DOI: 10.1109/IGARSS.2014.6946357.

Zhao, M., Running, S.W., Heinsch, F.A., Nemani, R.R. 2011. MODIS-derived terrestrial primary production. – Ramachandran, B., Justice, C.O., Abrams, M.J. (eds.). Land Remote Sensing and Global Environmental Change: NASA's Earth Observing System and the Science of ASTER and MODIS. New York, Springer-Verlag, 635–660.

Appendix A1. Soil type classification.
Lisa A1. Mullakaardi andmetest moodustatud klassid.

Table A1.1. The grouping schema of soils for machine learning.
Tabel A1.1. Mullakaardi andmestiku rühmitamise skeem masinõppe jaoks.

Soil / Muld		Soil / Muld		Soil / Muld		Soil / Muld	
Type / *Tüüp*	Code *Kood*	Type / *Tüüp*	Code *Kood*	Type / *Tüüp*	Code *Kood*	Type / *Tüüp*	Code *Kood*
Ag	48	GI1	45	L	61	LkII	51
AG	48	Gk	14	L(k)	61	LkII(g)	51
AG1	48	Gk1	21	L(k)g	63	LkIIg	53
AM′	16	Gkr	14	L(k)I	61	LkIII	51
AM″	37	Go	21	L(k)Ig	63	LkIII(g)	53
AM‴	48	Go1	45	L(k)II	61	LkIIIg	53
Ar	48	GoI	45	L(k)IIg	63	LP	42
ArG	48	Gor	21	L(k)III	61	LP(g)	43
Arv	109	Gr	14	L(k)IIIg	63	LPe	42
Av	109	K	11	LG	64	LPg	43
D	42	K(g)	11	Lg	73	LPG	44
D(g)	42	Kg	13	LG1	64	M	37
Dg	43	Kh	10	LI	61	M′	16
DG	44	Khg	13	LIg	73	M″	37
E2I	51	KI	31	LII	61	M‴	48
E2k	11	KI(g)	42	LIIg	73	Pp	45
E2o	31	KIg	42	LIII	61	R	77
E3I	51	Ko	21	LIIIg	73	R′	57
E3k	10	Ko(g)	21	Lk	51	R″	77
E3o	31	Kog	21	Lk(g)	53	R‴	77
G(o)	21	Kor	21	LkG	44	S	57
G1	45	Korg	21	Lkg	53	S′	57
Gh	14	Kr	10	LkI	51	S″	57
Gh1	14	Kr(g)	11	LkI(g)	51	S‴	77
GI	45	Krg	13	LkIg	53	-	-

Appendix A2. Confusion matrices of pixel level dominant tree species estimates.
Lisa A2. Enamuspuuliigi hinnangute veamaatriksid.

Table A2.1. Cross-tabulation of estimated (columns) and known (rows) main species code in National Forest Inventory (NFI) sample plots where dominant species proportion is more than 75%. User, producer and overall accuracies are presented.

*Tabel A2.1. Puistuplaanil (veerud) ja statistilise metsainventuuri (SMI) proovitükkidel teadaoleva (read) enamuspuuliigi risttabel vaatlustele, kus enamuspuuliigi koosseisukordaja üle 75%. Tabelis on toodud ka üldine-, kasutaja-, tootjatäpsus (*O.ACC, User ACC *ja* Prod. ACC*).*

NFI sample plots/ *SMI proovitükid*		Estimated main species / *Enamuspuuliiik*							Prod. ACC
		11	12	13	14	15	16	25	
Silver birch	11	259	12	20	29	6	17	2	0.75
Norway spruce	12	30	200	43	3	2	5	0	0.71
Scots pine	13	30	27	573	1	2	7	0	0.90
Gray alder	14	18	2	0	71	2	4	0	0.73
Black alder	15	22	1	3	7	25	5	1	0.39
European aspen	16	20	8	2	4	2	23	2	0.38
Other species	25	3	6	0	14	4	8	4	0.10
User ACC		0.68	0.78	0.89	0.55	0.58	0.33	0.44	O. ACC = 75.5%

Table A2.2. Cross-tabulation of estimated (columns) and known (rows) main species code in National Forest Inventory (NFI) sample plots where dominant species proportion is more than 75% and stands are older than 20 years. User, producer and overall accuracies are presented.

*Tabel A2.2. Puistuplaanil (veerud) ja statistilise metsainventuuri (SMI) proovitükkidel teadaoleva (read) enamuspuuliigi risttabel üle 20-aastastes puistutes, kus enamuspuuliigi koosseisukordaja üle 75%. Tabelis on toodud ka üldine-, kasutaja-, tootjatäpsus (*O.ACC, User ACC *ja* Prod. ACC*).*

NFI sample plots *SMI proovitükid*		Estimated main species / *Enamuspuuliik*							Prod. ACC
		11	12	13	14	15	16	25	
Silver birch	11	228	12	17	15	4	13	2	0.78
Norway spruce	12	17	186	42	0	1	5	0	0.74
Scots pine	13	20	25	552	1	2	4	0	0.91
Gray alder	14	17	1	0	49	2	2	0	0.69
Black alder	15	20	1	3	5	22	4	1	0.39
European aspen	16	18	8	2	2	2	20	1	0.38
Other species	25	3	5	0	9	2	5	4	0.14
User ACC		0.71	0.78	0.90	0.60	0.63	0.38	0.50	O. ACC = 78.4%

Appendix A3. Map of main dominant species in Estonia.

Lisa A3. Eesti puistute enamuspuuliigi kaart.

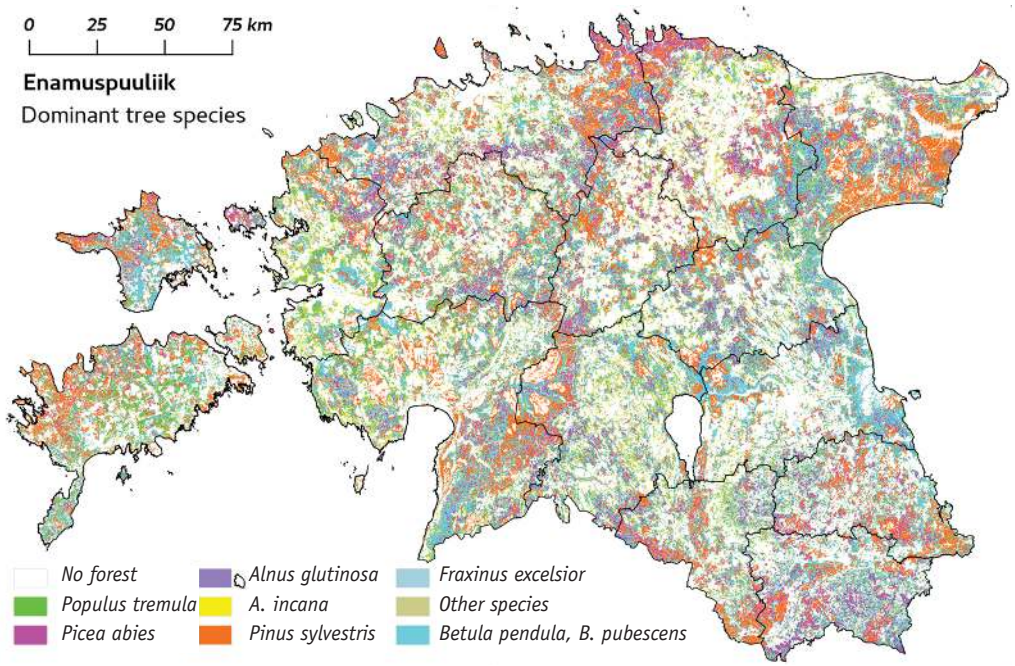

Figure A3.1. Dominant species of the estimated species composition. County borders are imposed upon the map. Colour figure is available in electronic version.

Joonis A3.1. Puistute liigilise koosseisu enamuspuuliigi kaart. Kaardile on lisatud maakondade piirid. Värviline kaart on e-ajakirjas.

Eesti puistute liigilise koosseisu hindamine multispektraalsete satelliidipiltide, mullakaardi ja näidistel põhineva masinõppe abil

Mait Lang, Mihkel Kaha, Diana Laarmann ja Allan Sims

Kokkuvõte

Jätkusuutliku metsamajanduse otsuste kavandamine vajab metsaökosüsteeme kirjeldavaid andmed, mis sõltuvalt kavandamise eesmärgist võivad olla erineva ruumilise, temaatilise ja ajalise üldistustasemega. Puistupõhised takseerandmed saadakse metsakorralduse käigus (Metsakorralduse, 2017), riigi tasemel ülevaade saadakse statistilise metsainventuuri abil (Tomppo *et al*., 2010). Samas on puistupõhiste takseerandmete uuendamine metsaomaniku jaoks enamasti vajalik teatud pikema perioodi järel või siis ainult neis metsades, kus soovitakse kavandada majandusotsuseid. Statistilise metsainventuuri vaatluste põhjal saab üldistusi teha kõige rohkem maakonna tasandil. Teisalt kasvab nõudlus pidevalt ülepinnaliselt uuendatavate andmete järele uut tüüpi interaktiivse metsakorralduse süsteemide arenedes (Korjus *et al*., 2017), seoses vajadusega korraldada seiret või rakendada taimkatte primaarproduktsiooni mudeleid (Zhao *et al*., 2011; Lang *et al*., 2017). Ühe lahendusena saab kasutada kogu ala katvaid kirjeldavaid tunnuseid multispektraalsetelt satelliidipiltidelt ning mõnda näidistel põhinevat masinõppemeetodit, mille realisatsioone on testitud ka Eestis (Tamm & Remm, 2009).

Käesoleva uuringu eesmärgiks oli koostada Eestit kattev puistute koosseisu hinnang keskmise ruumilahutusega (25 m) digitaalse kaardina. Selleks kasutati *random forest* masinõppemeetodit (Breiman 2001) GRASS GIS paketist r.learn.ml (GRASS Development Team, 2017). Kogu ala kirjeldavate tunnuste jaoks saadi Landsat-8 OLI ja Sentinel-2 MSI pildid (tabel 1) USGS GloVis (https://glovis.usgs.gov) ja Copernicus Open Access Hub (https://scihub.copernicus.eu) pildiarhiividest ning 1:10 000 digitaalne mullakaart (Mullakaardi, 2001) Eesti Maa-ameti kodulehelt. Puistute näidiseid võeti metsaregistri andmebaasi (Forest database, 2016) koopiast seisuga 13.02. 2017. Satelliidipildid teisendati Eesti põhikaardi koordinaatsüsteemi, valides kompromissina piksli suuruseks 25 m. Pilve- ja pilvevarjude maskid digiti käsitsi. Metsaregistri andmebaasist võeti päringuga välja alates 2014. aastast takseeritud metsamaa eraldised suurusega 1–10 ha. Esmasesse valimisse sattunud näidiseraldistel puhverdati piirid 30 m sissepoole ning siis jäeti alles ainult need, mis sisaldasid vähemalt 16 pikslit. Kirjeldavate tunnuste väärtuste arvutamiseks näidistele kasutati eraldise polügoni tsentroidi ümber 36 m raadiuses leiduvaid piksleid. Algset näidiste valimit puhastati vigadest puistute heleduse ja vanuse ning heleduse ja tüvemahu seoste järgi ning peale teede ja kraavide läheduses olevate näidiste kõrvaldamist jäid masinõppe jaoks alles 102 291 puistu andmed.

Mullakaardi andmed üldistati tabeli A1.1 järgi. Masinõpet rakendati alampiirkondade kaupa (joonis 1) järgmiste sammudena iga pildi või pildipaari (fenoloogilise info kasutamiseks) korral vastavalt:

1. informatiivsete tunnuste otsing;
2. õpetusandmete ettevalmistamine;
3. *random forest* algoritmi hüperparameetrite optimumi otsing;
4. sihtpikslitele puistu koosseisu ennustamine.

Töös kasutatud *random forest* meetodi realisatsioon ennustas igale pikslile kõikide klasside (puuliigid ning arenguklass lage/selguseta) tõenäosused. Eeldati, et puistu

spektraalne signatuur satelliidipildi pikslitel on käsitletav koosseisuliikide spektraalsete signatuuride lineaarkombinatsioonina ja saadud tõenäosusi tõlgendati puuliikide osakaaludena puistu liigilises koosseisus.

Tulemuste valideerimiseks kasutati puistute pindala jagunemist enamuspuuliigi järgi aastaraamatust Mets (Raudsaar *et al.*, 2017) (joonis 2); uuendusraietel kogutud harvestermõõtmise andmeid Riigimetsa Majandamise Keskuse andmebaasist (tabel 2, joonis 3); metsa kasvukäigu püsiproovitükkide andmebaasi (Kiviste *et al.*, 2015) väljavõtet (tabel 3, joonis 4) ja statistilise metsainventuuri (SMI) proovitükke (tabel 4, joonis 5). Valideerimise tulemusena selgus, et maakonna suuruse ala keskmisena on puistuplaani enamuspuuliigi pikslite jaotus väga heas kooskõlas statistiliste andmetega nii riigi ($R^2 = 0.98$) kui erametsade ($R^2 = 0.93$) osas. Puistute ja proovitükkide tasemel tehtud vaatluste keskmistatud tulemuste põhjal ilmneb aga hinnangutes vähemlevinud puuliikide osakaalude ülehindamine ja enamlevinud puuliikide osakaalu allahindamine koosseisus. Enamuspuuliiki prognoositakse 78,4%-lise täpsusega puistutes, kus domineeriva liigi osakaal on vähemalt 75% ning puistu on üle 20 aasta vana (tabel A2.2). Üsna hästi hinnatakse okaspuuliikide osakaalu koosseisus, kuid lehtpuuliikide omavahelise eristamise osas oleks vaja edasisi uuringuid. Esineb ka mõningane tendents hinnangu vigade suurenemisele nooremates puistutes (joonis 5, tabelid A2.1 ja A2.2). Üheks vigade põhjuseks on asjaolu, et alati olid ennustustes esindatud kõik klassid ning juba paariprotsendine juhusliku esinemise tõenäosus iga klassi jaoks annab summaarselt üheksa klassi kokkuvõttes üsna märkimisväärse vea. Selle tõttu puhtpuistud tulemuseks saadud liigilises koosseisu hinnangutes praktiliselt puuduvad. Maakonniti tehtud valideerimine ja okaspuude osakaalu analüüs ning koosseisu osakaalude regressioonanalüüs näitasid, et käesolevas töös masinõppe meetodil saadud kaart on kasutatav puistuplaanina (joonis A3.1), mis võimaldab saada puistute liigilise koosseisu hinnangu kogu Eesti ulatuses aladel, kus ajakohased metsa takseerandmed puuduvad.

Received March 6, 2018, revised May 25, 2018, accepted July 30, 2018