Mapping growing stock volume and forest live biomass: a case study of the Polissya region of Ukraine

Environmental Research Letters

LETTER

Mapping growing stock volume and forest live biomass: a case study of the Polissya region of Ukraine

Andrii Bilous[1], **Viktor Myroniuk**[1], **Dmytrii Holiaka**[1], **Svitlana Bilous**[1], **Linda See**[2] and **Dmitry Schepaschenko**[2,3]

[1] National University of Life and Environmental Sciences of Ukraine, Heroyiv Oborony 15, 03041, Kyiv, Ukraine
[2] International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria
[3] Author to whom any correspondence should be addressed.

E-mail: schepd@iiasa.ac.at

Keywords: data fusion, *k*-NN imputation, random forest, model-based inference, confidence interval

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
28 March 2017

REVISED
9 July 2017

ACCEPTED FOR PUBLICATION
1 August 2017

PUBLISHED
27 September 2017

Abstract

Forest inventory and biomass mapping are important tasks that require inputs from multiple data sources. In this paper we implement two methods for the Ukrainian region of Polissya: random forest (RF) for tree species prediction and *k*-nearest neighbors (*k*-NN) for growing stock volume and biomass mapping. We examined the suitability of the five-band RapidEye satellite image to predict the distribution of six tree species. The accuracy of RF is quite high: ~99% for forest/non-forest mask and 89% for tree species prediction. Our results demonstrate that inclusion of elevation as a predictor variable in the RF model improved the performance of tree species classification. We evaluated different distance metrics for the *k*-NN method, including Euclidean or Mahalanobis distance, most similar neighbor (MSN), gradient nearest neighbor, and independent component analysis. The MSN with the four nearest neighbors ($k = 4$) is the most precise (according to the root-mean-square deviation) for predicting forest attributes across the study area. The *k*-NN method allowed us to estimate growing stock volume with an accuracy of 3 m^3 ha^{-1} and for live biomass of about 2 t ha^{-1} over the study area.

1. Introduction

Data from remote sensing (RS) are crucial for a number of tasks in forest inventory and monitoring, including the estimation of forest area and its dynamics, construction of thematic forest maps, estimation of tree species distribution, stratification of the territory for sampling, and calculation of forest parameters for an area from a sample (Latifi *et al* 2015a, McRoberts *et al* 2014, Schepaschenko *et al* 2015a Schepaschenko *et al* 2015b). The challenges of the combined implementation of ground-based and remote methods of forest inventory have been discussed extensively in the scientific literature (Chirici *et al* 2016, McRoberts and Tomppo 2007). In this regard, the technology of satellite image processing by the *k*-nearest neighbors (*k*-NN) method and random forest (RF) have been quite successful.

According to McRoberts (2012), the *k*-NN method is used in forest inventory for four main tasks: (1) imputation of missing values for forest inventory and forest monitoring databases; (2) wall-to-wall mapping based on point measurements; (3) small area estimation; and (4) support for design-based and model-based inference. Forest parameter estimation based on the *k*-NN technique involves combining sample plot measurements, RS data, and auxiliary information, including a digital elevation model (DEM), land cover, etc. Given the limited extent of ground measurements, the method provides reasonable accuracy, both overall and at a fine scale. The *k*-NN technique builds a continuous, spatially explicit model of each forest parameter. The model reflects the distribution and variability of forest indicators and results in a set of maps that support forest monitoring and inventory (Beaudoin *et al* 2014, Maselli and Chiesi 2006, McRoberts and Tomppo 2007, Mozgeris 2008, Reese *et al* 2003, Tomppo *et al* 2016).

Non-parametric methods are common for classification of satellite imagery. They do not have specific requirements for the distribution of the studied parameters. RF is one of the most efficient and

well-established machine learning techniques for classification of remote sensing imagery (Breiman 2001, Belgiu and Drăguț 2016) and can also be used for regression analysis. It relies on bagging to form an ensemble of classification and regression trees. Bootstrap samples are used to construct multiple trees such that each tree is split with a randomized subset of features, thus the name 'random' forest. Another subset, which is not part of the classification, is used for accuracy assessment. Recently, the method was successfully applied for deriving forest parameters (Latifi *et al* 2012) and tree species prediction in particular (Immitzer *et al* 2012, Myklush *et al* 2013).

Although both the *k*-NN and RF methods have been commonly used in forestry applications as outlined in Chirici *et al* (2016), a meta-analysis by the same authors revealed that only 3.4% of the 148 papers reported confidence intervals while most were limited to overall accuracy, the Kappa index and root-mean-square deviation (RMSD) (see e.g. Bernier *et al* 2011, Gagliasso *et al* 2014, Latifi *et al* 2015b, Trubins and Sallnäs 2014, Zald *et al* 2016). In contrast this paper reports the estimated forest parameters with standard errors, which are needed for forest inventories. The forest mask produced using RF is also provided with confidence intervals and compared with other forest masks.

This paper discusses the implementation of both methods: (1) RF for delineating forested area and classifying tree species; and (2) *k*-NN for mapping growing stock volume and live biomass. We report confidence intervals for the estimated parameters using an area in Ukraine as a case study.

The traditional Ukrainian forest inventory and planning (FIP) approach involves first dividing the forest (using RS data) into homogenous units (primary units of forest inventory) and then forest parameters (i.e. tree species composition, age, average height and diameter, growing stock volume, etc.) are assigned by trained personal on the ground through visual estimation or measurements. This results in the description of the individual forest stands with a return interval of about 10 years. The limitations include: (1) FIP covers area managed by the state forestry authority only (about 85% of forested land); (2) shelterbelts and other protective tree associations in agroforestry systems are not covered; (3) independent monitoring is needed. Reliable remote sensing techniques are especially relevant for Ukraine, as about 15% of the forested area is not covered by the FIP and also because of the relatively large dynamics of forest cover, some of the drivers of which are illegal logging, dieback of shelterbelts due to drought, and afforestation of abandoned arable land.

2. Input datasets

2.1. Study area

The study area is in the Snovsk district of the Chernihiv region (figure 1) and covers 45 km^2. It is located within the East European Plain with wavy plan relief typical of the Ukrainian Polissya. The major part of the area belongs to the first and second river terraces with the altitude ranging between 120 and 170 m above mean sea level. According to the FIP completed in 2011, forest cover represents 1893 ha (463 individual forest stands) or 41.8% of the study area. Pine and birch are the most common species (44.7% and 39.8% of the forest area, respectively), alders cover 13.1%, while other species (aspen, oak, ash, black locust, spruce, and maple) each cover less than 1%. Forests are highly productive (reaching an average stand height of 27–34 m at 100 years old); the age structure has two peaks: young (37%) or mature (42%) forests are dominant, while the middle aged, premature, and over-mature groups are less well represented.

The study area is representative of the entire Ukrainian Polissya region in terms of landscape, tree species composition and productivity.

2.2. Spatial datasets

We used five-band RapidEye imagery with a spatial resolution of 5 m, acquired in 2011. The image was both geometrically corrected and converted to the top of atmosphere (TOA) reflectance. Another input dataset was a 10 m DEM resampled to the same resolution. All datasets were converted to a WGS 84/UTM zone 36 N projection (figure 2).

An IKONOS image with a spatial resolution of 1 m was used for visual interpretation of land cover type in order to validate the forest mask.

2.3. Live biomass model

The FIP database consists of information for every individual forest stand. This includes tree species and several other parameters, which were used to estimate the live biomass of all the forest stands as follows (Shvidenko *et al* 2007):

$$R_{fr} = {}^{M_{fr}}/_{\mathrm{GS}} = a_0 + A^{a_1} \cdot \mathrm{SI}^{a_2} \cdot \mathrm{RS}^{a_3} \cdot \exp(a_4 \cdot A + a_5 \cdot \mathrm{RS}) \quad (1)$$

where R_{fr} is the biomass expansion factor; M_{fr} is the live biomass of fraction *fr*, oven-dry t ha^{-1}; GS is the growing stock volume in m^3; A is age in years; RS is relative stocking; SI is the site index, which reflects the quality of the site (Shvidenko *et al* 2007); and a_0–a_5 are model parameters.

The biomass expansion factors (table S1, available at stacks.iop.org/ERL/12/105001/mmedia) for birch, black alder and aspen were estimated using 443 sample plots collected in Ukraine and neighboring countries, where the biomass fractions were measured using the destructive sampling method (Schepaschenko *et al* 2017b). Parameters for the pine and oak forests were taken from Shvidenko *et al* (2007) and applicable for European Russia, Belarus and Ukraine. Table S1 also reports the RMSD, which is the square root of the

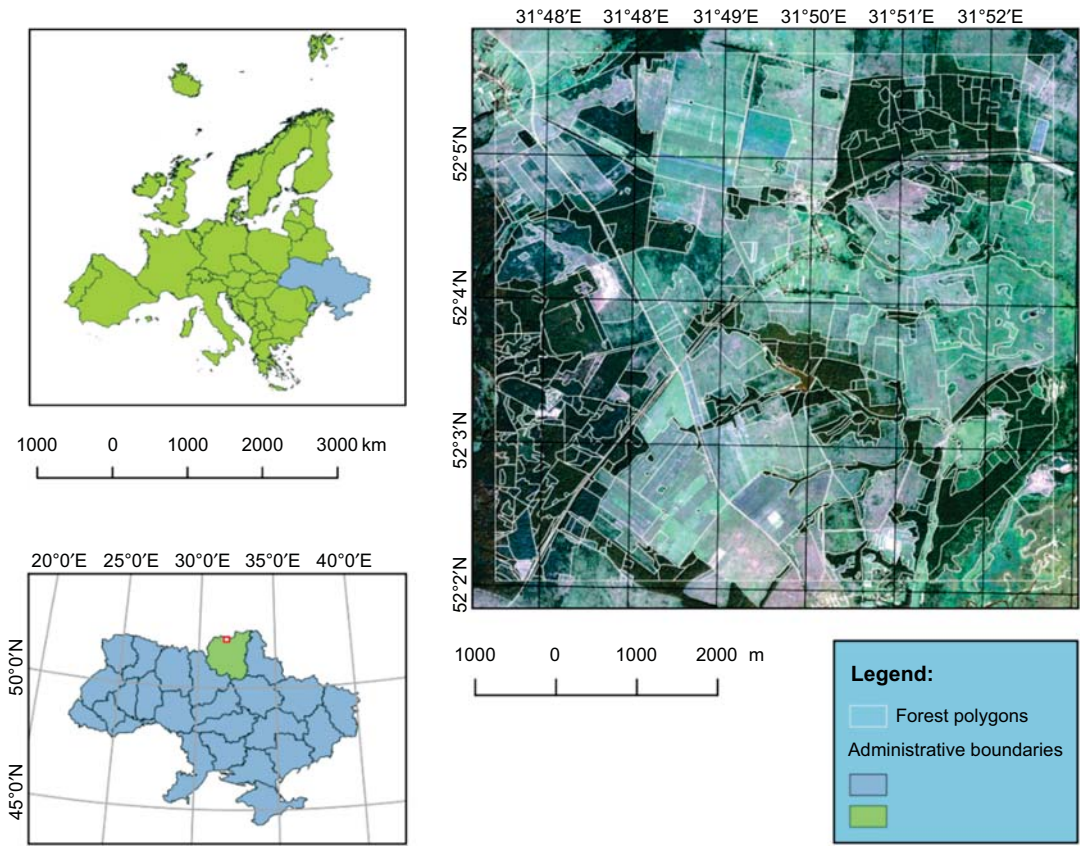


Figure 1. Location of the study area in the northern part of Ukraine.

Table 1. Dominant tree species and size of training dataset for RF classification.

Tree species	Latin name	ID	Sample size, points
Alder	*Alnus glutinosa* L	ALGL	188
Birch	*Betula pendula* Roth	BEPE	530
Pine	*Pinus sylvestris* L	PISY	854
Aspen	*Populus tremula* L	POTR	72
Oak	*Quercus robur* L	QURO	55
Black locust	*Robinia pseudoacacia* L	ROPS	30

average of the squared deviations between the actual and estimated values of the live biomass.

The live forest biomass of every forest stand was estimated based on equation (1) and the parameters presented in table S1. The models considered stem wood biomass over bark, aboveground live biomass of the stand (stem wood over bark, branches and leaves), the total live biomass of forest stands (aboveground stand biomass and roots) and the total forest live biomass (forest stand, understory and green forest floor) in an oven-dry state.

3. Methods

3.1. Random forest

The RF technique was used (1) to create a forest mask and (2) to delineate the dominant tree species. In our analysis, we used five-band *RapidEye* imagery at a 5 m spatial resolution and a DEM. The reference data were taken from the FIP database in the following way. We randomly distributed 4000 points over the study area. Of these, 1729 fell within the forest and species information from the FIP database, which were then used as a training dataset for classification. The main tree species and sample sizes are presented in table 1.

The results of the RF classification were aggregated in order to remove small groups of pixels (under 40). Finally, the minimum size of a forest polygon was 0.1 ha, which corresponds to the national regulations for the forest inventory.

The validation dataset consists of an additional 2300 randomly selected points. We calculated the confusion matrix and estimated the user's and producer's accuracies and the confidence intervals for the area of different tree species assessment (Congalton and Green 2008, Olofsson *et al* 2014).

The data were classified in *R* using the {randomForest} package, v. 4.6–12 (Liaw and Wiener 2002). To understand the contribution of each variable in the classification model, we used the mean decrease in accuracy (MDA) as a measure of the variable's importance. The

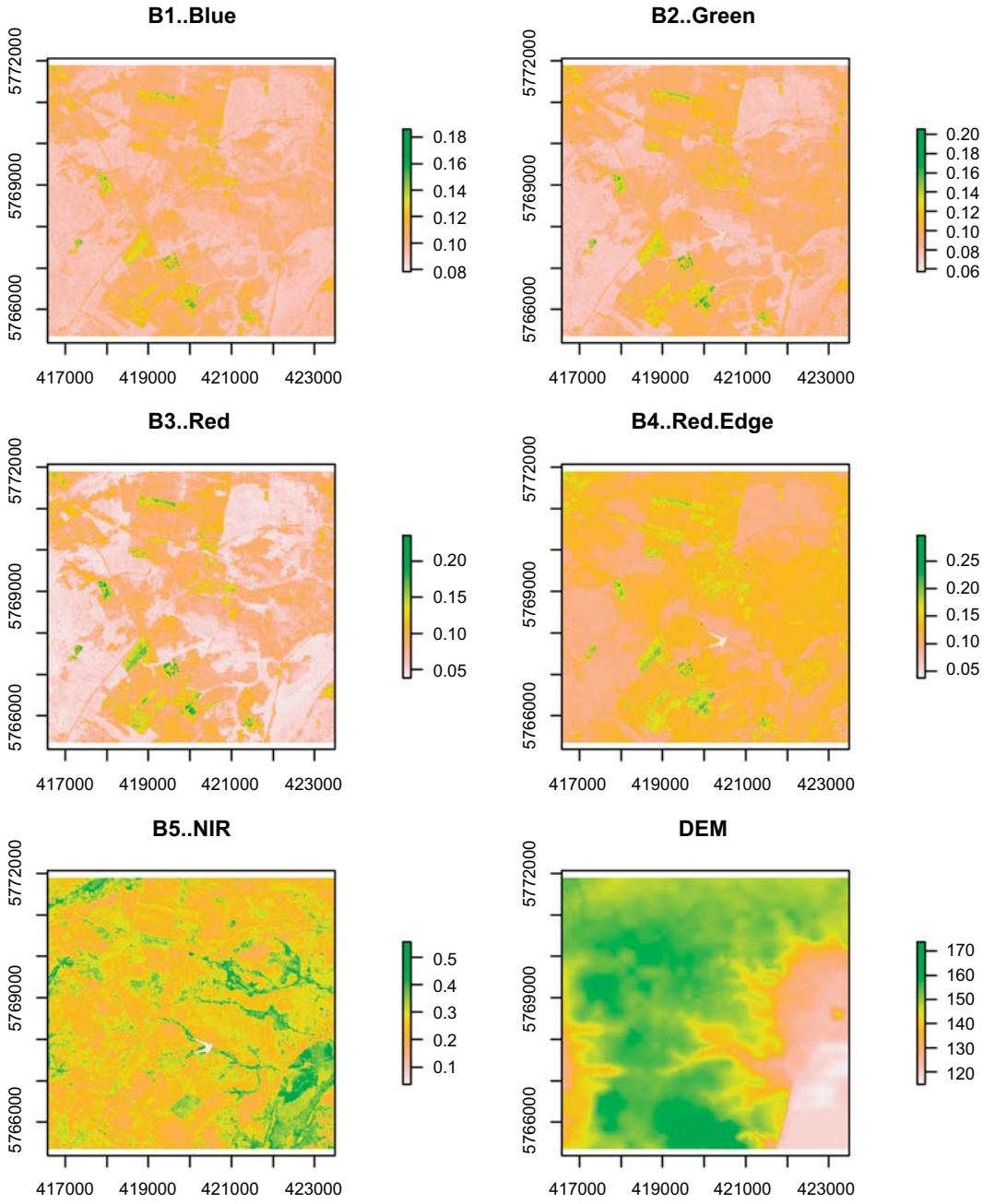


Figure 2. The examples of datasets used as independent variables in the forest classification: B1–B5—RapidEye bands; DEM—elevation above sea level, m.

general idea of the MDA is to rank variables according to their contribution (in percentage terms) to the mean squared error of a classification if they are excluded from the calculation. Clearly, the larger the MDA, the more the variable contributes to the accuracy of the model.

3.2. *k*-NN imputation

Imputation is a process that replaces missing values with predicted or observed values (McRoberts 2009). The *k*-NN technique was used to map growing stock volume and forest biomass. Ground truth information was obtained from the FIP database and consists of 90 randomly selected forest stands.

The *k*-NN method requires that both the number (k) of nearest neighbors and the equation to calculate the distance to these neighbors in the parameter space be specified. We compared several methods to estimate the distance, as suggested by Crookston and Finley (2008), namely, the Euclidean (EUC) and Mahalanobis (MAL) distances between the reflectance of the target and the reference pixels of the image. Other methods of calculating distance are based on canonical correlation analysis (MSN—most similar neighbor), canonical correspondence analysis (GNN—gradient nearest neighbor), and component analysis (ICA—independent component analysis).

The response variable (growing stock volume or live biomass) for a pixel p ($\tilde{y}_p$) was predicted by equation (2) (Tomppo *et al* 2016):

$$\tilde{y}_p = \sum_{i \in I_h} w_{i,p} y_i \quad (2)$$

where y_i is the value of variable y of the i-th member of the training dataset for pixel p; $w_{i,p}$ is the weight of the i-th member for pixel p; and I_h is the size of training sub-dataset for the stratum h.

The weight, $w_{i,p}$, is inversely proportional to the distance (d^t) between the target pixel and the NN in the parameter space as follows:

$$W_{i,p} = \frac{1}{d^t_{p_{i,p}}} \Big/ \sum_{j \in \{i_1(p),\ldots i_k(p)\}} \frac{1}{d^t_{p_{j,p}}} \quad (3)$$

where k is the number of nearest 'neighbors'; and t is a real number usually between 0 and 2. We use $t = 2$ to increase the contribution of large distances.

Besides classification and mapping, the k-NN method can be successfully used for the unbiased estimation of the mean values of the forest parameters based on a sample (McRoberts 2012). Model-based methods are used extensively in forest inventories to infer the mean values of the forest attributes, where the estimates are required for a small area of the population (small area estimates). If we assume that a simple model describes the population, then the expected values of Y at an i-th point belonging to the area of interest can be estimated as follows:

$$y_i = \mu + \varepsilon_i, \quad (4)$$

where μ is the mean value of Y for the population unit; and ε_i is the random deviation of observation y_i from its mean value μ at a point i.

Model-based approaches are generally focused on the estimation of the mean values of forest attributes in the population rather than on definite observations. With the k-NN method equation (2), the estimator of the population mean is $\hat{\mu}$:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \tilde{y}_p, \quad (5)$$

where N is the sample size.

The k-NN model was run in *R* using the {yaImpute} package, version 1.0–26 (Crookston and Finley 2008).

3.3. Estimation of the performance of global/regional products in the study area

We compared a number of global and regional maps with the ground FIP data in the study area in order to estimate their accuracy and if locally calibrated models are able to improve this accuracy further. The FIP data were aggregated to match the pixel size of each map. We predicted the forested area, the average and total biomass, the overall accuracy of the forest mask and the tree species group. Note that the estimated accuracies are only valid for the study area and cannot be treated as an estimation of uncertainties of the global/regional maps.

4. Results and discussion

4.1. Forest mask

The contribution of different parameters to the forest extent prediction is shown in figure 3. The most important input datasets for delineating the forest are the B5 NIR band, the coordinates and the elevation.

The very high resolution image, with the forest mask overlaid (revealing the edges of the imagery so that forest areas can be clearly seen), is presented in figure 4 with correspondence evident. The misclassification rate was estimated at 1.6% based on the OOB (out-of-bag) error.

The forest mask was used for further segregation of the forest area by dominant tree species.

We compared the forest extent in several global and regional maps of high resolution (25–60 m) with a random sample of visual interpretation of the IKONOS image (table 2). We estimated the user's and producer's accuracy of the forest masks and confidence interval for area estimation. Two of the maps (Hansen *et al* 2013, Sexton *et al* 2013) are represented as percentage tree cover. We compared the percentage of forest on the IKONOS image with the tree cover value on the maps (table 2, forest share RMSD). To delineate forest in these two maps, we applied a threshold (table 2, tree cover threshold) that matched the forested area as closely as possible without decreasing the accuracy.

The global land cover and forest maps delineate forest with reasonably good user's accuracy (75%–90%). GlobeLand30 shows the least amount of forest, which results in the lowest producer's accuracy. Despite the high resolution of this product, i.e. 30 m, it recognizes core forest areas only, classifying the rest as cropland, i.e. the dominant land cover type. Only the national dataset (Lesiv *et al* 2015) demonstrates high and balanced user's and producer's accuracies with narrow confidence intervals despite underestimating the forest area.

4.2. Tree species classification

The DEM, together with B5 NIR, contribute the most to the accuracy of the model (figure 5). Elevation was particularly important for the prediction of alder forest, which covers the lowest elevations, i.e. floodplains.

The OOB error (in our case 8.22%) shows a sufficiently high accuracy of prediction for the tree species.

Figure 6 shows the variability of independent variables used for tree species classification. Despite the fact that an individual variable may have overlapping values for different species, each contributes to a robust tree species prediction. For example, pine

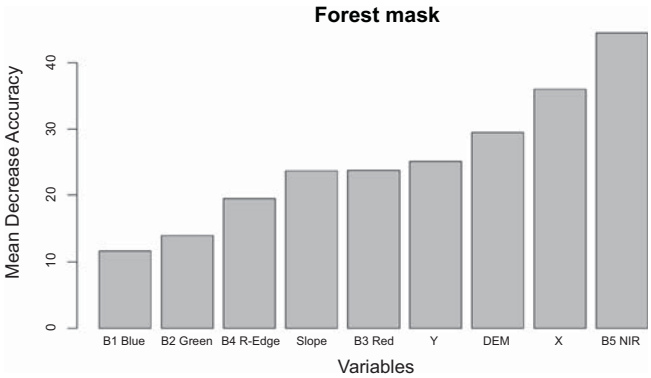


Figure 3. Importance of the variables for predicting the forest mask using a RF classification model.

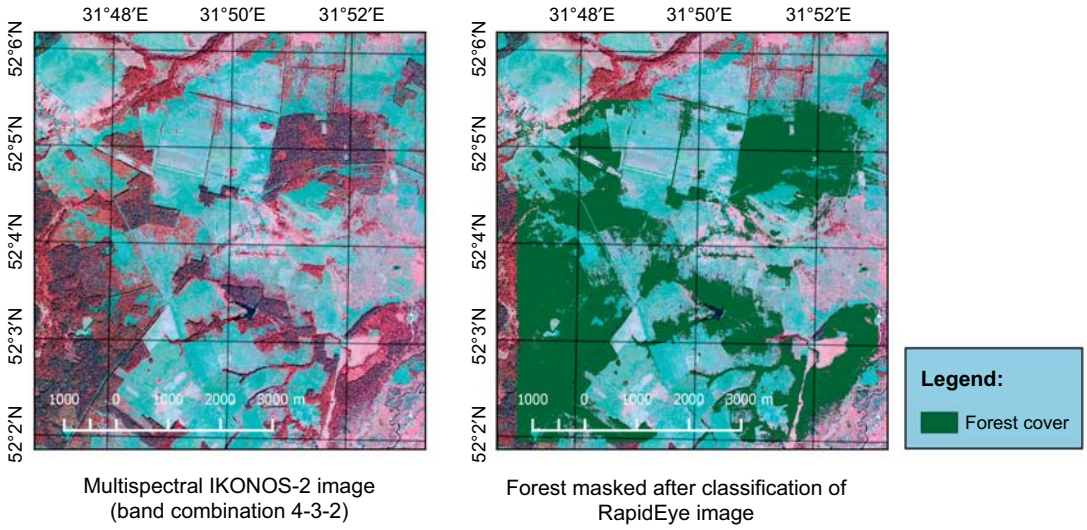


Figure 4. High resolution IKONOS image (left) and forest mask (green color) overlay the image (right).

Table 2. Performance of different global and national datasets in the study area.

Dataset	Pixel size (m)	Forest share RMSD (%)	Tree cover threshold (%)	Forested area on the map (%)	User's accuracy (%)	Producer's accuracy (%)	Adjusted forested area (%)	CI (%)
Our forest mask	5 × 5	–	–	35.2	99	97	36.6	0.8
PALSAR forest mask (Shimada *et al* 2014)	15 × 25	–	–	38.1	76	81	35.7	2.2
GlobeLand30 (Jun *et al* 2014)	19 × 30	–	–	16.3	90	41	35.7	2.6
Global tree cover (Hansen *et al* 2013)	19 × 30	22.6	25	32.2	86	77	35.2	2.0
Landsat VCF (Sexton *et al* 2013)	19 × 30	20.5	30	40.8	75	85	35.9	2.2
Ukrainian forest (Lesiv *et al* 2015)	38 × 60	–	–	31.4	92	81	35.7	1.8

forests are distinguished by both low near-infrared radiation (NIR) and red-edge values.

As expected, the spectral reflectance of the tree species has large values in the NIR and red-edge channels of RapidEye imagery, which means that these two parameters contribute the most in tree species recognition. The median value of the pine reflectance is significantly lower compared to all deciduous species while alder is the most distinguishable species among the deciduous trees based on these spectral channels.

The results of the classification are presented in the form of a tree species map for the study area (figure 7). The map was used for further quantification of growing stock values and live biomass.

From figure 7, one can see that the territory is quite heterogeneous in terms of species composition. A significant part of the forest area is covered with a pine (PISY, 48.7%), birch (BEPE, 27.8%) and alder (ALGL, 17.7%) dominated forest. Low elevation in the south-eastern part of the study area is covered by alder, while oak forest is observed in the western

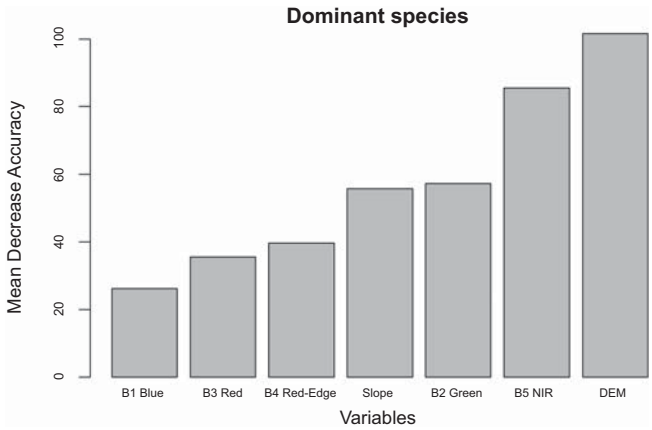


Figure 5. Importance of the variables for predicting the tree species using a RF classification model.

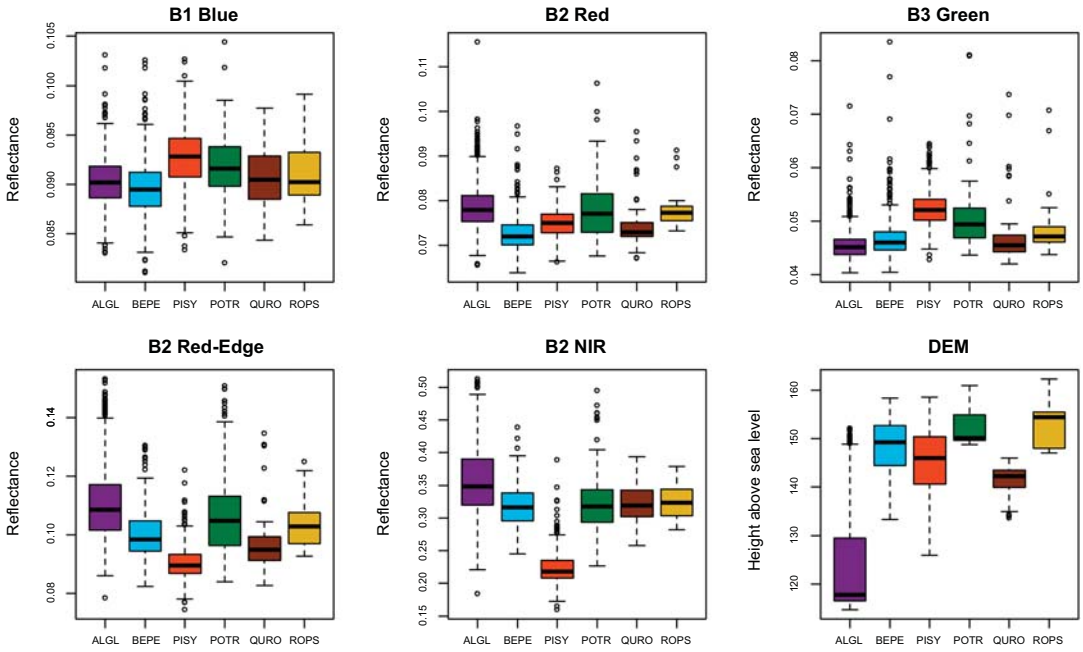


Figure 6. Box-whisker plots of median reflectance of the five RapidEye bands and DEM values for the six tree species.

part only. In general, it is consistent with the forest inventory data. The results of the validation of the tree species map based on 2300 randomly selected pixels are presented in table 3. The RF model is more accurate for alder and pine, which are represented by a pure (single species) stand in the study area. Birch is, in most cases, mixed with other species, and therefore birch-dominated forests are less accurate (table 3).

The overall accuracy of the tree species classification is 87.9%. The relatively low value of the producer's accuracy for oak, black locust, and aspen can be attributed to the small portion of the overall area covered by these species. For instance, oak is represented by five forest stands with an overall area of 3 ha while there are only two 3.5 ha black locust stands. On the other hand, the user's accuracy overall (except for oak) exceeds 75%. The most robust model in terms of both user's and producer's accuracies was obtained for pine and alder, where sufficient training data were available for the RF algorithm.

We checked how global and regional datasets describe tree species in the study area. We indicate the share of forest area recognized by every map (table 4) and the accuracy of classifying coniferous evergreen and broadleaved deciduous species. Even at the tree group level, the error exceeded 25% and even 50% in many cases. The Global Land Cover by National Mapping Organizations (GLCNMO) classified 1/3 of the forested area as 'broadleaf evergreen' or 'needleleaf deciduous', which are not represented in the study area. MODIS Land Cover classifies all the forest as mixed, which is more or less correct at a 500 m resolution.

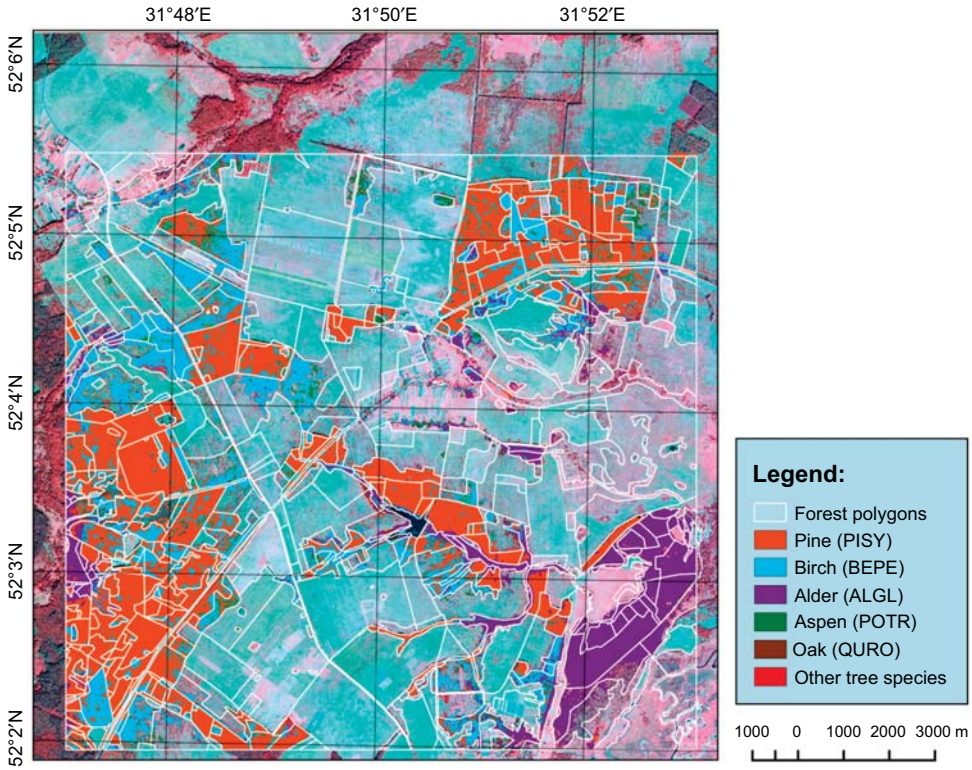


Figure 7. The dominant tree species predicted using the RF model.

Table 3. Confusion matrix of tree species classification.

Classified	References						Total	Producer's accuracy	User's accuracy	Area, ha	SE, ha ($p = 0.95$)
	ROPS	BEPE	ALGL	QURO	POTR	PISY					
ROPS	0.61	0	0	0	0.14	0.07	0.8	75.0	46.8	21	7
BEPE	0.60	20.29	2.45	0.95	1.80	1.70	27.8	73.0	93.4	347	18
ALGL	0.04	0.43	16.75	0.14	0.20	0.12	17.7	94.8	83.4	320	13
QURO	0.05	0.21	0.11	0.74	0	0.21	1.3	56.0	36.5	32	9
POTR	0	0.23	0.15	0	2.77	0.54	3.7	75.0	50.6	87	13
PISY	0	0.57	0.64	0.19	0.57	46.75	48.7	96.0	94.7	787	16
Total	1.3	21.7	20.1	2.0	5.5	49.4	100	—	—	1594	—

Table 4. Performance of different global and local datasets in delineating coniferous evergreen and broadleaved deciduous species.

Dataset	Pixel size (m)	Forest area recognized (%)	Overall accuracy (%)
Our forest mask	5 × 5	84	95.4
Ukrainian forest (Lesiv *et al* 2015)	38 × 60	81	70.5
ESA CCI LC (ESA Land Cover CCI project team and Defourny P 2016)	190 × 310	55	75.2
European tree species (Brus *et al* 2012)	1000 × 1000	no mask	35.7
GLCNMO (Tateishi *et al* 2014)	285 × 463	47	20.2

4.3. Growing stock volume and biomass

Optical imagery can be used for the indirect estimation of biomass based on canopy cover and surface reflectance (Schepaschenko *et al* 2017a, Lu *et al* 2016). We tested several *k*-NN imputation methods to predict GSV and LB (table 5).

Table 5 demonstrates the advantage of the MSN method, which has the smallest normalized RMSD for every response variable. This method was thus selected for further implementation. To determine the optimal number of *k* for use with the NN technique, we started with a small number and then increased the number iteratively until the accuracy no long increased significantly. Figure 8 shows that the highest accuracy was reached at $k = 4$. More NNs provide the same or a larger normalized RMSD.

As a result of the *k*-NN imputation, we obtained GSV and LB maps at a spatial resolution of 5 m (figure 9). The pine forest has the largest GSV (325 m^3 ha^{-1} on average). There is minor variability in the GSV within stocked forest stands. Sparse stands, which are mainly encroaching on former arable land, have much higher variability of GSV. High variability is observed for small biomass forests

Table 5. Normalized RMSD for different *k*-NN imputation methods for growing stock volume and live biomass estimation.

	Normalized RMSD for different imputation methods				
Response variable	EUC	MAL	MSN	ICA	GNN
Growing stock volume	0.7456	0.7477	0.5936	0.7477	0.7581
Stem biomass	0.7785	0.7600	0.6380	0.7600	0.8041
Above-ground biomass (AGB)	0.7869	0.7644	0.6518	0.7644	0.8082
Tree biomass (above- and below-ground)	0.8273	0.7943	0.6895	0.7943	0.8309
Forest biomass	0.8311	0.7961	0.6917	0.7961	0.8333

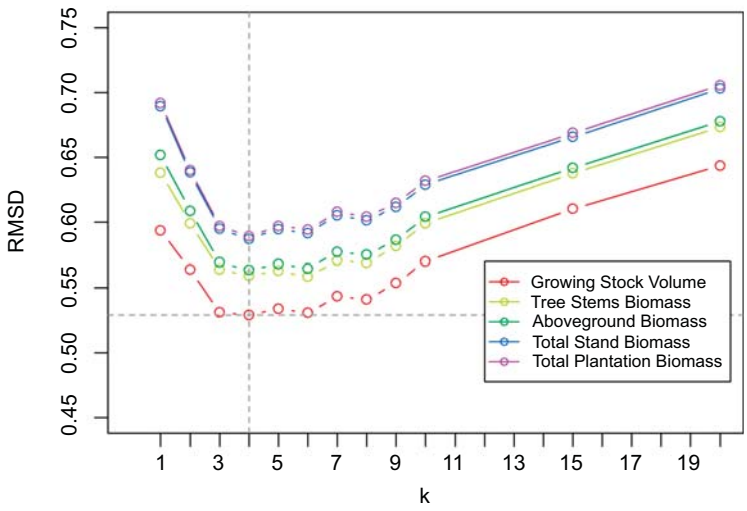


Figure 8. Normalized RMSD of growing stock volumes and biomass for MSN imputation and different number of neighbors (*k*).

(up to 75 m^3 ha^{-1}) and biomass is highly correlated with GSV (figure 9).

We used equation (5) to estimate the mean growing stock volume and the live biomass for the study region (table 6).

The average modeled GSV over the study area is 165 m^3 ha^{-1}, which corresponds to forest inventory data of 162 m^3 ha^{-1}. Pine has the largest GSV and biomass values per hectare.

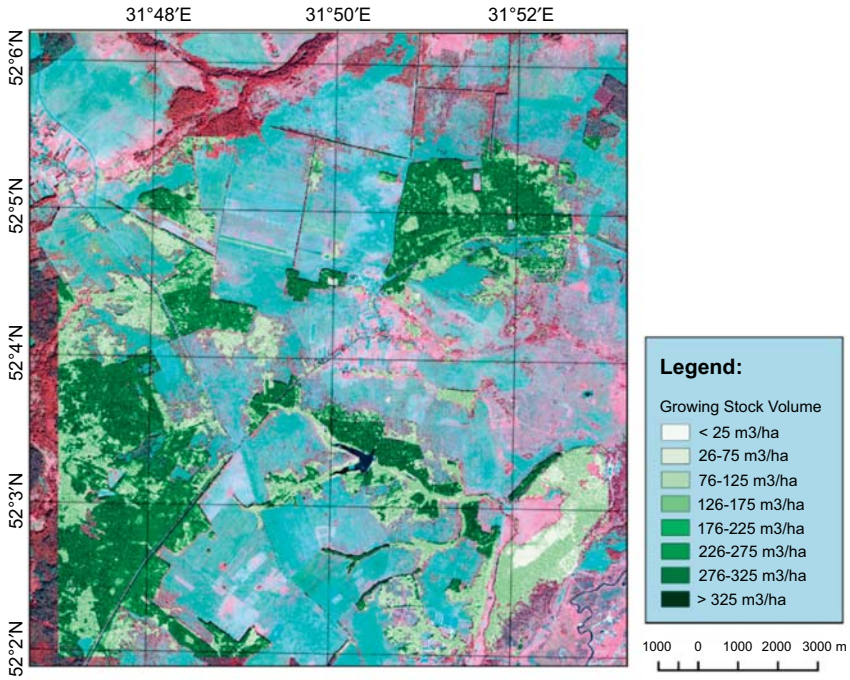
The distribution of the imputed GSV and biomass in the study area for different tree species and aggregated for all species together is presented in figure 10. The distribution has two peaks. The second one corresponds to mature pine forests with large GSV and biomass.

We compared the biomass from different global and regional maps (table 7). Most of the datasets have a 1 km resolution so they are too coarse for calculating the accuracy of the spatial distribution of biomass over the study area. Hence we only compared the average biomass value for the forest area recognized by the datasets and the total biomass for the study area. The average AGB estimates vary from −31% up to +49% compared to the FIP data. The most similar estimates were obtained for the national map produced by Lesiv *et al* (2015) and the boreal biomass map by Thurner *et al* (2014). All the datasets substantially underestimated the total biomass (from −7% to −72%). This is the result of both underestimation of forested area and the average biomass value.

5. Conclusions

The integration of the information derived from satellite images, a DEM, other geospatial datasets, and a limited number of field measurements can contribute to the effective prediction of forest attributes at the pixel level during a forest inventory. The application of RF and *k*-NN techniques allows for an unbiased estimation of the mean values of forest parameters and the mapping of the forest based on remote sensing data with a small number of ground measurements. Both methods are viable for the processing of RapidEye images for area estimation, prediction of tree species composition, and imputation of structural forest parameters. In the context of the remote sensing assisted estimation of the growing stock volume and the live biomass, our research demonstrates how forests can be mapped in the form of continuous surfaces and how the mean values of forest attributes can be assessed across a defined geographic region.

The methods applied here demonstrate a substantial increase in accuracy compared to existing global and national products. Global forest masks capture forests well in the study area, but existing tree species distributions and biomass estimations are poor. Hence the proposed methods are very promising for national-scale forest inventories and monitoring in Ukraine, especially considering illegal harvesting, the dieback of shelterbelts due to droughts, and the afforestation of abandoned arable land.



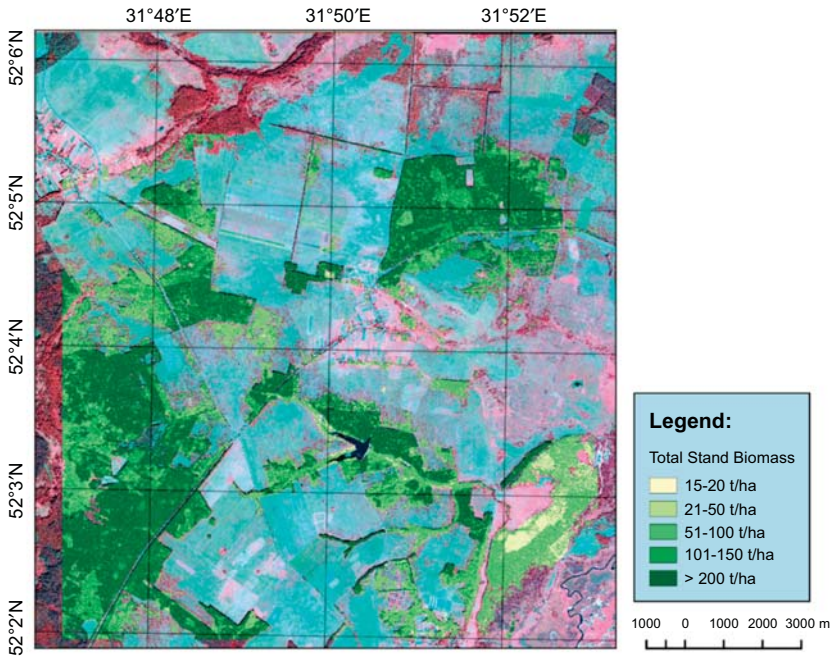


Figure 9. Predicted growing stock volume and total stand biomass over the study area using *k*-NN imputation (MSN, $k = 4$) with an IKONOS image in background.

Table 6. Mean growing stock volume and biomass in forest stands according to the *k*-NN method.

		Biomass ($t\,ha^{-1}$)			
Tree species	Growing stock volume ($m^3\,ha^{-1}$)	stem	above-ground	total stand	total forest
BEPE	129	64	74	101	104
ALGL	125	55	64	83	86
QURO	141	61	72	93	96
POTR	184	79	94	120	125
PISY	313	139	156	188	193
Other sp.	100	46	54	72	75
Mean by *k*-NN	**165**	**74**	**86**	**109**	**113**
Mean by FIP[a]	**162**	**75**	**84**	**109**	**112**

[a] Forest inventory and planning weighted average data for 463 forest stands with total area of 1893 ha.

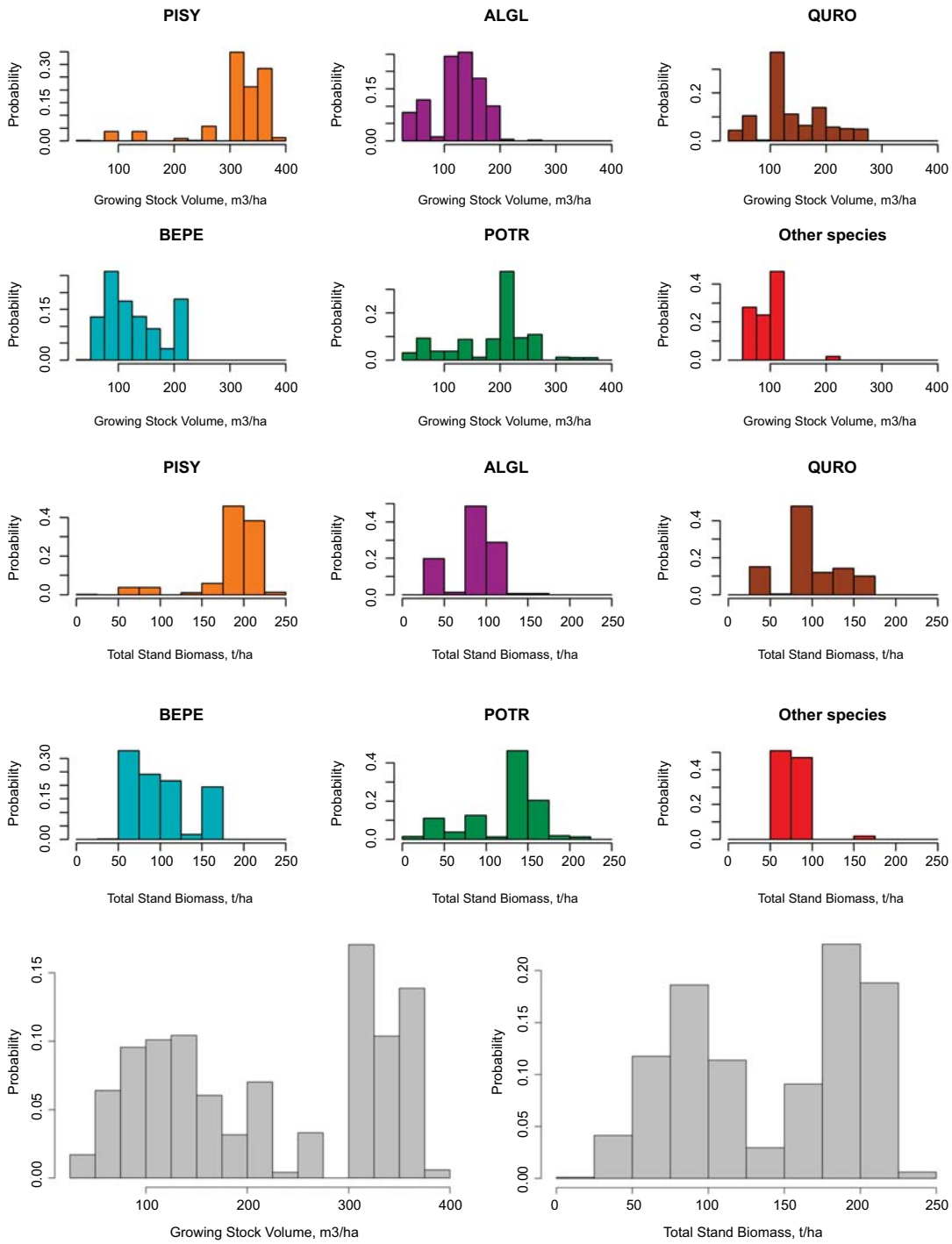


Figure 10. Distribution probabilities for growing stock volume and biomass in forests according to the *k*-NN technique.

Table 7. Comparison between different above ground tree biomass (AGB) estimates.

Dataset	Average AGB of forested area (t ha^{-1})	Total AGB (10^3 t)
Forest inventory	84	133
Our forest biomass map	86	134
European map by Gallaun *et al* (2010)	97	117
Boreal by Thurner *et al* (2014)	79	36
European by Kindermann *et al* (2008)	58	124
Global by Kindermann *et al* (2008)	125	114
Global by Hu *et al* (2016)	100	37
Ukraine by Lesiv *et al* (2015)	80	120

Acknowledgments

This work was partly support by the Ministry of Education and Science of Ukraine as well as by the European Space Agency under the DUE GlobBiomass project (contract 4000113100/14/I-NB).

ORCID iDS

Viktor Myroniuk https://orcid.org/0000-0002-5961-300X
Linda See https://orcid.org/0000-0002-2665-7065
Dmitry Schepaschenko https://orcid.org/0000-0002-7814-4990

References

Beaudoin A, Bernier P Y, Guindon L, Villemaire P, Guo X J, Stinson G, Bergeron T, Magnussen S and Hall R J 2014 Mapping attributes of Canada's forests at moderate resolution through *k*-NN and MODIS imagery *Can. J. Forest Res.* **44** 521–32

Belgiu M and Drăguţ L 2016 Random forest in remote sensing: a review of applications and future directions *ISPRS J. Photogramm. Remote Sens.* **114** 24–31

Bernier P Y, Daigle G, Rivest L-P, Ung C-H, Labbé F, Bergeron C and Patry A 2010 From plots to landscape: a *k*-NN-based method for estimating stand-level merchantable volume in the province of Québec, Canada *Forest Chron.* **86** 461–8

Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

Brus D J, Hengeveld G M, Walvoort D J J, Goedhart P W, Heidema A H, Nabuurs G J and Gunia K 2012 Statistical mapping of tree species over Europe *Eur. J. Forest Res.* **131** 145–57

Chirici G, Mura M, McInerney D, Py N, Tomppo E O, Waser L T, Travaglini D and McRoberts R E 2016 A meta-analysis and review of the literature on the *k*-nearest neighbors technique for forestry applications that use remotely sensed data *Remote Sens. Environ.* **176** 282–94

Congalton R G and Green K 2008 *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (New York: CRC Press)

Crookston N L and Finley A O 2008 yaImpute: an R package for *k*-NN imputation *J. Stat. Softw.* **23** 1–16

ESA Land Cover CCI project team and Defourny P 2016 ESA land cover climate change initiative (land cover CCI) dataset collection *Centre for Environmental Data Analysis, 2017* (http://catalogue.ceda.ac.uk/uuid/c19b0914521144ab8c18c91d586c6847)

Gagliasso D, Hummel S and Temesgen H 2014 A comparison of selected parametric and non-paramentric imputation methods for estimating forest biomass and basal area *Open J. Forestry* **4** 42–48

Gallaun H, Zanchi G, Nabuurs G J, Hengeveld G, Schardt M and Verkerk P J 2010 EU-wide maps of growing stock and above-ground biomass in forests based on remote sensing and field measurements *Forest Ecol. Manage.* **260** 252–61

Hansen M C *et al* 2013 High-resolution global maps of 21st-century forest cover change *Science* **342** 850–3

Hu T, Su Y, Xue B, Liu J, Zhao X, Fang J and Guo Q 2016 Mapping global forest aboveground biomass with spaceborne lidar, optical imagery, and forest inventory data *Remote Sens.* **8** 565

Immitzer M, Atzberger C and Koukal T 2012 Tree species classification with random forest using very high spatial resolution 8-band worldview-2 satellite data *Remote Sens.* **4** 2661–93

Jun C, Ban Y and Li S 2014 China: open access to Earth land-cover map *Nature* **514** 434–4

Kindermann G E, McCallum I, Fritz S and Obersteiner M 2008 A global forest growing stock, biomass and carbon map based on FAO statistics *Silva Fenn.* **42** 387–96

Latifi H, Fassnacht F E, Hartig F, Berger C, Hernández J, Corvalán P and Koch B 2015a Stratified aboveground forest biomass estimation by remote sensing data *Int. J. Appl. Earth Obs. Geoinf.* **38** 229–41

Latifi H, Fassnacht F E, Hartig F, Berger C, Hernández J, Corvalán P and Koch B 2015b Stratified aboveground forest biomass estimation by remote sensing data *Int. J. Appl. Earth Obs. Geoinf.* **38** 229–41

Latifi H, Nothdurft A, Straub C and Koch B 2012 Modelling stratified forest attributes using optical/LiDAR features in a central European landscape *Int. J. Digit. Earth* **5** 106–32

Lesiv M, Shvidenko A, Schepaschenko D, See L and Fritz S 2015 Forest map and its uncertainty as an important input for carbon sink estimation for Poland and Ukraine *Proc. 4th International Workshop on Uncertainty in Atmospheric Emissions, 7–9 October 2015* (Krakow: Systems Research Institute, Polish Academy of Sciences) pp 9–15

Liaw A and Wiener M 2002 Classification and regression by random forest *R News* **2** 18–22

Lu D, Chen Q, Wang G, Liu L, Li G and Moran E 2016 A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems *Int. J. Digit. Earth* **9** 63–105

Maselli F and Chiesi M 2006 Evaluation of statistical methods to estimate forest volume in a Mediterranean region *IEEE Trans. Geosci. Remote Sens.* **44** 2239–50

McRoberts R E 2009 Diagnostic tools for nearest neighbors techniques when used with satellite imagery *Remote Sens. Environ.* **113** 489–99

McRoberts R E 2012 Estimating forest attribute parameters for small areas using nearest neighbors techniques *Forest Ecol. Manage.* **272** 3–12

McRoberts R E, Liknes G C and Domke G M 2014 Using a remote sensing-based, percent tree cover map to enhance forest inventory estimation *Forest Ecol. Manage.* **331** 12–8

McRoberts R E and Tomppo E O 2007 Remote sensing support for national forest inventories *Remote Sens. Environ.* **110** 412–9

Mozgeris G 2008 Estimation and use of continuous surfaces of forest parameters options for Lithuanian forest inventory *Balt. Forest* **14** 176–84

Myklush S I, Chaskovskyy O H and Gavrylyuk S A 2013 Remote sensing data dectyption for the assessment of tree species groups *Proc. Forest Acad. Sci. Ukr.* **11** 144–55

Olofsson P, Foody G M, Herold M, Stehman S V, Woodcock C E and Wulder M A 2014 Good practices for estimating area and assessing accuracy of land change *Remote Sens. Environ.* **148** 42–57

Reese H, Nilsson M, Pahén T G, Hagner O, Joyce S, Tingelöf U, Egberth M and Olsson H 2003 Countrywide estimates of forest variables using satellite data and field data from the national forest inventory *Ambio* **32** 542–8

Schepaschenko D *et al* 2015a Development of a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics *Remote Sens. Environ.* **162** 208–20

Schepaschenko D *et al* 2017a Forest biomass observation: current state and prospective *Sib. J. Forest Sci.* **4** 3–10

Schepaschenko D *et al* 2017b A dataset of forest biomass structure for Eurasia *Sci. Data* **4** sdata201770

Schepaschenko D, Shvidenko A Z, Lesiv M Y, Ontikov P V, Shchepashchenko M V and Kraxner F 2015b Estimation of forest area and its dynamics in Russia based on synthesis of remote sensing products *Contemp. Probl. Ecol.* **8** 811–7

Sexton J O *et al* 2013 Global, 30 m resolution continuous fields of tree cover: landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error *Int. J. Digit. Earth* **6** 427–48

Shimada M, Itoh T, Motooka T, Watanabe M, Shiraishi T, Thapa R and Lucas R 2014 New global forest/non-forest maps from ALOS PALSAR data 2007–2010 *Remote Sens. Environ.* **155** 13–31

Shvidenko A, Schepaschenko D, Nilsson S and Bouloui Y 2007 Semi-empirical models for assessing biological productivity of Northern Eurasian forests *Ecol. Model.* **204** 163–79

Tateishi R, Hoan N T, Kobayashi T, Alsaaideh B, Tana G and Phong D X 2014 Production of global land cover data—GLCNMO 2008 *J. Geogr. Geol.* **6** 99–122

Thurner M *et al* 2014 Carbon stock and density of northern boreal and temperate forests *Glob. Ecol. Biogeogr.* **23** 297–310

Tomppo E, Haakana M, Katila M and Peräsaari J 2008 *Multi-Source National Forest Inventory* vol 18 (Dordrecht: Springer Netherlands)

Trubins R and Sallnäs O 2014 Categorical mapping from estimates of continuous forest attributes–classification and accuracy *Silva Fenn.* **48** 1–16

Zald H S J, Wulder M A, White J C, Hilker T, Hermosilla T, Hobart G W and Coops N C 2016 Integrating landsat pixel composites and change metrics with lidar plots to predictively map forest structure and aboveground biomass in Saskatchewan, Canada *Remote Sens. Environ.* **176** 188–201