Research paper

An indirect approach to predict deadwood biomass in forests of Ukrainian Polissya using Landsat images and terrestrial data

Maksym Matsala*, Viktor Myroniuk, Andrii Bilous, Andrii Terentiev, Petro Diachuk and Roman Zadorozhniuk

Matsala, M., Myroniuk, V., Bilous, A., Terentiev, A., Diachuk, P., Zadorozhniuk, R. 2020. An indirect approach to predict deadwood biomass in forests of Ukrainian Polissya using Landsat images and terrestrial data. – Forestry Studies | Metsanduslikud Uurimused 73, 107–124, ISSN 1406-9954. Journal homepage: <http://mi.emu.ee/forestry.studies>

Abstract. Spatially explicit and consistent mapping of forest biomass is one of the key tasks towards full and appropriate accounting of carbon budgets and productivity potentials at different scales. Landsat imagery coupled with terrestrial-based data and processed using modern machine learning techniques is a suitable data source for mapping of forest components such as deadwood. Using relationships between deadwood biomass and growing stock volume, here we indirectly map this ecosystem compartment within the study area in northern Ukraine. Several machine learning techniques were applied: Random Forest (RF) for the land cover and tree species classification task, *k*-Nearest Neighbours (*k*-NN) and Gradient Boosting Machines (GBM) for the deadwood imputation purpose. Land cover (81.9%) and tree species classification (78.9%) were performed with a relatively high level of overall accuracy. Outputs of deadwood biomass mapping using *k*-NN and GBM matched quite well ($8.4 \pm 2.3 \text{ t ha}^{-1}$ (17% of the mean) vs. $8.1 \pm 1.7 \text{ t ha}^{-1}$ (16% of the mean), respectively mean $\pm$ SD deadwood biomass stock), indicating a strong potential of ensemble boosters to predict forest biomass in a spatially explicit manner. The main challenges met in the study were related to the limitations of available ground-based data, thus showing the need for national statistical inventory implications in Ukraine.

Key words: deadwood mapping, Random Forest, *k*-Nearest Neighbours, gradient boosting, biomass.

Authors' address: Department of Forest Mensuration and Forest Management, National University of Life and Environmental Sciences of Ukraine, Heroiv Oborony Street 15, 03041, Kyiv, Ukraine; *matsala@nubip.edu.ua

Introduction

Spatially explicit estimation of forest structural compartments has become a main target of studies and forest monitoring during the last decades (Lim *et al.*, 2003; Zimble *et al.*, 2003; Zald *et al.*, 2016). Deadwood as the key legacy component of forest ecosys-

tems reflecting natural dynamics processes in the stand structure, is increasingly considered in conservation and management policies throughout the world (Franklin *et al.*, 2002; Seidl *et al.*, 2014). Consistent assessment of deadwood stocks has been recognized as an important task in the studies of biodiversity and habitat maintenance,

DOI: 10.2478/fsmu-2020-0018

© 2020 by the authors. Licensee Estonian University of Life Sciences, Tartu, Estonia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

carbon sequestration, nutrient cycling, pest controlling and other ecosystem functions and services (Angelstam & Kuuluvainen, 2004).

Landsat image time series with simultaneous pre-processing of spectral data that are freely available on the cloud-computing platform Google Earth Engine have created new capabilities for mapping of forests and their biophysical parameters and features (Cohen & Goward, 2004). New temporally continuous and spatially explicit approaches have significantly empowered a robust estimation of deadwood (Gonzalez *et al.*, 2018). Temporal segmentation of Landsat seasonal mosaics allows an identification of natural disturbances which are strictly linked to deadwood accumulation (Cohen *et al.*, 2010; Kennedy *et al.*, 2010). Other remote-based methods have been developed (e.g., Oeser *et al.*, 2017), focused on different natural drivers of tree mortality: either stand-replacing events (e.g., wildfires, intensive storms) or gap-scale disturbances (e.g., insect outbreaks). Significant improvement of deadwood estimation has been achieved with rising application of terrestrial-based and airborne LiDAR techniques coupled with empirical field-based and satellite-based data (Gonzalez *et al.*, 2018).

Herewith, precise estimation of forested areas and quantification of ecosystems' biophysical values is linked to the various statistical classification and regression approaches. For the last decades, non-parametric machine learning methods have shown a strong capability to be applied for forest remote sensing data. A number of studies refer to successful implementation of numerous classifiers (e.g., decision trees (Ruefenacht *et al.*, 2008); support vector machines (SVM, Li *et al.*, 2013); artificial neural networks (Civco, 1993)) with application to coarse and medium (Nitze *et al.*, 2015), high (Thanh Noi & Kappas, 2018) and extra-high (Hayes *et al.*, 2014) spatial resolution of data. These methods have shown good performance in classify-

ing land cover types (Gislason *et al.*, 2006), forest successional stages (Liu *et al.*, 2008), and tree species occupation (Evans & Cushman, 2009). Regression tasks like estimation of growing stock volume (GSV), live and dead biomass and sequestered biomass carbon have met more robust and user-friendly implementation since R-package "yaImpute" and similar applications were developed (Crookston & Finley, 2008; McRoberts, 2012).

To date, ensemble machine learning methods are successfully used in remote-sensing applications (Belgiu & Dragut, 2016). Using a set of so-called weak learners (mainly decision trees), these models apply either a bootstrap or bagging process (well-known Random Forest, RF), or boosting fitting (e.g., gradient boosting machines, GBM). The main challenges those ensemble and above-mentioned methods face are the quality of reference and ancillary data, while tuning of modelling hyper-parameters (e.g., learning rate of number of iterations) was reported as an issue more dependent on dataset features (Maxwell *et al.*, 2018).

In the twenty-first century, Ukraine is facing both an increasing interest in a full carbon accounting of forest ecosystems, including estimation of deadwood stocks, and substantial challenges due to the absence of the main sources of reference data (e.g., sampling plots-based National Forest Inventory, LiDAR applications, Shvidenko *et al.*, 2017). Polissya is the second forested region in Ukraine and can be described by specific circumstances that call for robust applications of deadwood estimation. Those are the presence of primeval forests of black alder (*Alnus glutinosa* L.), recovery processes on abandoned areas due to the radioactive contamination since the Chornobyl disaster in 1986 and economic shifts, and land degradation caused by illegal amber mining (Lesiv *et al.*, 2018). Some studies refer to the possible assessment of deadwood stocks using field sample plots and multispectral satellite images in this region

(Bilous *et al.*, 2017; Lakyda *et al.*, 2019).

Here we aim to (i) examine the performance of creating forest mask using Random Forest for the study site in Ukrainian Polissya; (ii) map stocks of coarse woody debris (CWD) biomass using *k*-Nearest Neighbours (*k*-NN) and gradient boosting models (GBM).

Materials and Methods

Study area

A study site in Ukrainian Polissya was chosen with the aim to fully capture regional geographical and forest composition features. The study was carried out for a site in the Chernihiv region (Northern Ukraine, Figure 1) with an area of *ca* 1,192 km², between N 50°54'1", E30°44'60" and N51°14'15", E31°12'16". The dominated land cover types (according to the local operational forest management data) are croplands, grasslands in Desna river bank and homogeneous coniferous replanted forests (with Scots pine (*Pinus sylvestris* L.) as the main tree species). Some forest stands are formed by silver birch (*Betula pendula* Roth.) and common aspen (*Populus tremula* L.); natural forests of black alder are also presented. Pine-dominated forests are mainly pure, with insignificant admixtures (i.e., the share of mean GSV is *ca* less than 10%) of silver birch. Forests within the study area are managed by two main authorities: State Forest Enterprise "Osterskiy" and Regional Landscape Park "Mezhyrichensky".

The local climate represents the continental climate subtype, with a mean annual temperature of 7.2 °C and an average total precipitation of 702 mm. Soils are mainly sod-podzolic, with alluvial soils in river banks. Considering the forest type classification implemented in Ukraine, Scots pine grows there in A₂ and B₂ types (i.e., poor and semi-poor level of soil fertility and fresh sites in terms of humidity), silver birch – B₂ and C₃, and black alder –

C₄ and C₃ (soils with a rich nutrient content and high humidity level; for more detailed classification of forest types see Bilous *et al.*, 2019).

Data on deadwood biomass

Forest inventory maps delineating forest polygons have been incorporated in the study as a source of forest attributes derived from the forest planning and management (FPAM) data (available for 2011). We considered utilizing FPAM data instead of statistical forest inventory datasets, since these are currently limited to several regions (Myroniuk, 2018) which do not include the study area. FPAM data for these stands contain information on dominant tree species, mean diameter at breast height, mean height, age, site index, relative stocking, forest type, and GSV estimated during the last forest accounting. In total, we used data from 2,125 primary units available from the FPAM dataset for the study area. These polygons were given as ESRI shapefiles.

Stocks of CWD biomass were estimated using allometric linear models developed for the Ukrainian Polissya region (Lakyda *et al.*, 2019). According to the classification implemented in Ukraine (Bilous *et al.*, 2019), CWD is divided into such well-common deadwood components as snags (i.e., dead standing trees), logs (i.e., downed stems and stumps) and litter of coarse (*d* > 1 cm) branches. In this study, only biomass stocks of snags and logs were estimated: according to Lakyda *et al.* (2019), the litter of coarse branches weakly correlates with spectral data and mainly depends on the soil and fine litter features. The sum of the studied stocks further is considered as deadwood biomass stock. The allometric model used is given as (1):

$$M = a_1 \cdot D^{a_2} \cdot H^{a_3} \cdot S^{a_4} \quad (1)$$

where *M* is the biomass stock of snags or logs (t ha⁻¹), *a*₁, *a*₂, *a*₃ and *a*₄ are regression parameters, *D* is the diameter at breast

Figure 1. Location of the study site. The Chernihiv region (green) and study area (red) are coloured on the context map in the top. White lines refer to the edges of local forest management polygons used in this study, while the background is a raw Landsat image given in the band combination Near Infrared-Red-Green.

height (cm), H is the mean height (m), S is the relative stocking (ranging from 0 to 1). According to the source of these models (Lakyda *et al.*, 2019), there was a relatively small root mean square error (3% of the estimated mean). This error was ignored in

the following modelling and uncertainties estimations. For more detailed information on the regression parameters and predictor sets across compartments and tree species, see the respective paper (Lakyda *et al.*, 2019).

Remote sensing data

Accuracy and reliability of produced maps is strictly dependent on the agreement between the used terrestrial-based and remote sensing data. Considering that FPAM data for the study area was available only for 2011, the same point of time was chosen for obtaining remote sensing data. We used the open United States Geological Service archive presented in the Google Earth Engine (GEE, Gorelick *et al.*, 2017) to download Landsat-5 Thematic Mapper (TM) images (with 30 m spatial resolution) and create a cloud-free mosaic for the given site and time frame from 22 May 2011 to 22 September 2011 (according to the duration of the vegetation season in Ukraine). Landsat-5 TM imagery was already geometrically corrected. Radiometric correction to the surface reflectance and mosaicking with the aim of masking pixels with clouds were conducted using the GEE. The mosaics were then

downloaded from the GEE and processed in the R environment (R Core Team, 2019).

Geospatial modelling: classification and data imputation

The main workflow of this study is illustrated in Figure 2. Further, the methods of classification and data imputation are described as given in this flowchart. For the land cover classification task, the RF classifier was chosen as the most popular among the models used in the various scientific papers (e.g., Jhonnerie *et al.*, 2015; Thanh Noi & Kappas, 2018). Then, this type of a model was utilized for the tree species classification task (e.g., Lang *et al.*, 2018). The relatively novel approach of gradient boosting (e.g., Sachdeva *et al.*, 2018) was compared to the *k*-NN, as the last method is commonly used for regression and imputation purposes (e.g. Lang *et al.*, 2014; Latifi *et al.*, 2015; Fu *et al.*, 2019).

Figure 2. Workflow of the study. The flowchart illustrates utilization of the data, and the modelling process for classification and imputation purposes.

Classification reference data

We provided a two-step classification of Landsat mosaics. First, land cover classes were predicted using the RF machine learning algorithm to produce a forest cover mask. Then, the classification on dominant tree species across forest stands within the produced mask was performed. For both steps we used the same set of 6 predictors: Normalized Vegetation Difference Index (NDVI), Normalized Water Difference Index (NDWI), Tasseled Cap Transformation (TCT, namely brightness TCB, greenness TCG and wetness TCW), and elevation from the Digital Elevation Model (DEM) downloaded from the GEE for the study area. We chose NDVI and TCT as predictors because of their ability to capture GSV and biomass stock changes (Kauth & Thomas, 1976). In order to better classify water bodies, we utilized the NDWI predictor here. As an auxiliary non-spectral variable, elevation from the DEM is useful for distinguishing black alder from other broad-leaved species, since these stands are more likely to grow in lower sites regarding elevation above sea

level (Lakyda *et al.*, 2019). Ancillary datasets are illustrated in Figure 3.

Multispectral data for land cover classification was extracted for a set of 1,010 sample points, randomly distributed across the study area. Sample plots were generated in the Quantum Geoinformatics System (QGIS 3.4) software, and land cover classes were then visually interpreted using freely available software from Google Inc. (Google Earth for visual interpretation). Each sampling point was visually inspected using high-resolution images available on Google Earth as a reference to assign one of eight thematic classes: bog, croplands (including gardens), forests (including the stands of linear spatial shape in the fields, across roads and rivers), grasslands (including unforested sites surrounded by forests, but without signs of clear-cuts being identified with the time series of high-resolution images available on Google Earth), settlements, shrublands, water bodies (rivers and lakes) and other (sands, rocks).

The validation dataset for the land cover classification was also compiled inde-

Figure 3. Ancillary datasets used for spatial modelling. The DEM variable is given in meters above sea level.

pendently using the Collect Earth plugin and included 964 points, different from points used in the training dataset.

The dataset for tree species classification was extracted from the FPAM database for 742 points that fell within forested areas. All points were carefully interpreted on Google Earth to ensure they reliably represented characteristics of the inventory polygons which they were linked to. For the training model we used 557 points (75%) and the rest (185 points) were used for the validation task. We considered four dominating tree species in the study: PISY (acronym for Scots pine), BEPE (acronym for silver birch), ALGL (acronym for black alder), and POTR (acronym for common aspen). Other species were grouped in the thematic class OTHER.

Parameterizing classification models

RF is an ensemble method that decreases the variance of output, using a set of simple learners (here Classification and Regression Trees, CARTs) that preferably give outputs with low bias. The training dataset is split into small subsets used by each learner in

the RF model in an independent way, then the results are averaged (Breiman, 2001) to get the best performance (a process called “bootstrap”, or “bagging”) and the lowest variance. We used the “randomForest” R-package (Liaw & Wiener, 2002) to classify land cover. Default parameters (number of decision trees $n = 500$ and number of variables tried at each split by model $mtry = 3$) were reported as sufficient for good performance of the RF classifier, so we did not tune these (Belgiu & Dragut, 2016).

Performance of the RF models was assessed in terms of the following metrics: overall accuracy and the kappa (Congalton, 1991) index for all datasets, and user accuracy and producer accuracy for the class “forest”. Confusion matrixes were built in the “caret” R-package (Kuhn, 2008).

Imputation of deadwood biomass

We used a training dataset that consisted of 439 sample stands with information on deadwood biomass stocks calculated for each forest stand using the abovementioned allometric models. Such a number of stands and points was chosen because

Figure 4. Scatterplot of the relationship between mean GSV and deadwood biomass stocks within the study area. Coloured gradient represents the density of points.

of full spatial consistency between FPAM-based polygon map data and the forest mask obtained by the machine classifier for these stands. Imputation of continuous values of deadwood biomass stocks was performed using k -NN and GBM models based on the extracted mean spectral values of all pixels within the polygon of each stand from the sample dataset.

Since the allometric models of deadwood biomass are consistent with the parameters of a growing forest stand, we directly model this compartment instead of linking to GSV values (Pflugmacher *et al.*, 2012). This decision was made based on an apparently strong relationship between GSV and deadwood biomass values obtained from FMAP data for the study area (Figure 4).

Tuning imputation models

For imputation purposes, the performance of the k -NN algorithm was compared to GBM. The first method searches for an average value between the so-called “nearest neighbors” for a given point in the features’ space. The weighting of values of “neighbors” is defined by different schemes matrixes (McRoberts, 2012). In this study we used the “randomForest” method, which implies no weighting of matrixes in searching for the values of “nearest neighbors”, while the number of “neighbors” (k) was iteratively increased, with consideration of the threshold when the k -NN model likely starts to overfit. We iteratively increased the number of k (from 1 to 10), thus training 10 k -NN models. Imputation was performed in the R package “yalmpute” (Crookston & Finley, 2008).

GBM was implemented in the “gbm” (Greenwell *et al.*, 2019) R-package. Similar to the RF, it is an ensemble model, but weak learners (CART here as well as in the RF classifier) in it “boost” each other instead of performing independently like RF models. It means that in gradient boosting, every additional CART amplifies the performance of the previous learner

or is rejected if fails to provide more precise output. To tune the GBM model, it is necessary to control the various hyperparameters that define training velocity and performance on the datasets depending on their size. Here we built a hyperparameter grid (Table 1), thus running 540 different models in the algorithm that was searching for the lowest mean square error (MSE). Inner validation was provided by this tuning algorithm, which took for this purpose a random 25% subset for each model.

Table 1. Hyperparameter grid for tuning GBM models in this study.

Parameter to tune	Hyperparameter values				
Shrinkage	0.01	0.05	0.02	0.001	
Interaction depth	2	4	6	8	10
Number of trees	1000	5000	10000		
Number of observations in the terminal nodes	5	10	15		
Bagging fraction	0.5	0.75	1.0		

The shrinkage parameter (Table 1) defines a learning rate applied to each CART in the expansion. Interaction depth shows the highest level of variable interactions allowed in each tree. The number of trees shows how many CARTs were utilized in the GBM model. Bagging fraction is a proportion of observations randomly chosen to propose the next CART in the expansion, i.e., making the given GBM a stochastic boosting model (if this fraction is less than 1.0).

We assessed the produced GBM models, considering their root mean square error (RMSE) produced by a random inner validation subset (25% of the whole training dataset, as the training fraction for all models was 0.75). Performance of the GBM and k -NN models on the whole (without an inner validation subset) training dataset was compared in terms of RMSE, and visual analysis of fit against a 1:1 line.

Additionally, we compared model performance on a test landscape which was

not covered by the forest stands from the training dataset. This landscape (1,020 ha) represents forest communities typical of the entire study area: Scots-pine dominated stands with admixtures of silver birch and rare deciduous forests. Mean and total *k*-NN and GBM outputs within this landscape were compared to the FPAM data. Moreover, we estimated the mean net difference between reference and model data at the stand (polygon) scale.

Results

Land cover and tree species classification

The RF model produced has shown quite good overall accuracy (81.9%) and the kappa (74.6%) index. Land cover types “water”, “forest” and “croplands” on the validation dataset were interpreted with a high level of user accuracy (93.1%, 92.2%, and 90.0%, respectively), while the largest misclassifications were met for the types “bog”, “other” and “shrublands” (Table 2).

Table 2. Confusion matrix of RF land cover classification.

Predicted	Observed							
	Bog	Croplands	Forest	Grasslands	Others	Settlement	Shrubland	Water
Bog	13	0	1	0	0	0	1	0
Croplands	0	348	11	10	19	12	7	0
Forest	6	5	261	6	0	1	11	0
Grasslands	1	22	2	111	0	1	19	1
Others	0	0	0	0	0	1	0	0
Settlement	0	4	2	0	2	15	1	0
Shrubland	1	6	6	14	0	0	14	1
Water	1	0	0	0	0	0	0	27

The tree classification model has shown relatively good performance (overall accuracy is 78.9%, the kappa index is 56.4%), with user accuracy ranging from 90.0%

for Scots pine to 60.0% for silver birch and 80.0% for black alder. The class “OTHER” and common aspen achieved worse accuracies (23% and 29%, respectively).

Table 3. Confusion matrix of RF tree species classification.

Predicted	Observed				
	ALGL	BEPE	OTHER	PISY	POTR
ALGL	12	1	5	4	3
BEPE	0	12	3	7	0
OTHERS	0	0	3	1	2
PISY	3	7	2	117	0
POTR	0	0	0	1	2

Figure 5. Importance of predictors for specific tree species and in total given by the RF classification model.

Figure 6. Land cover classes map (a) and tree species map within the forest mask (b) produced by the respective RF model.

Figure 7. Agreement between predicted and observed values of deadwood biomass: imputed by the best k -NN (a, $k = 1$) and GBM (b) models.

Elevation was the most important variable to predict forests of black alder (Figure 5), while TCG and TCW contributed the most to the prediction of Scots pine stands. All predictors rather weakly classified less abundant broadleaved tree species (silver birch, common aspen and other), which resulted in higher errors (Table 3). Figure 6 illustrates the outputs of both RF classification models.

According to the map provided in Figure 6a, forest cover within the study area in 2011 was 40,367 ha, or 32.0% of the total area. This forested area was covered (Figure 6b) by stands of the following species (with 95% confidence interval, CI): 3320 ± 1328 ha of black alder forests ($8.2 \pm 3.3\%$), 3958 ± 1459 ha of silver birch forests ($9.8 \pm 3.6\%$), 1893 ± 1310 ha of common aspen forests ($4.7 \pm 3.2\%$), 27862 ± 1979 ha of Scots pine forests ($69.0 \pm 4.9\%$), and 3334 ± 1634 ha of stands where other species grew ($8.2 \pm 4.0\%$).

Imputation of deadwood biomass stocks

Tuning the GBM algorithm has shown that the lowest RMSE (2.09 t ha^{-1} , or 26% of the mean deadwood biomass stock) on the random 25% inner validation subset was achieved by a model with the following

hyperparameters: number of trees = 1,000, shrinkage = 0.01, minimum number of observations in the terminal node = 15, bagging fraction = 0.5, interaction depth = 10. The other nine best GBM models produced quite similar errors (ranged from 2.087 to 2.093 t ha^{-1}). According to their hyperparameters, the best learning rate for training the GBM in this case is quite high – 0.01. As well, all the best models used a bagging fraction value of 0.5 and a minimum number of observations in the terminal node of 15. Interaction depths there are mostly high (> 6), i.e., decision trees with a low number of “branches” did not produce the lowest error. In addition, models with various numbers of CARTs utilized were almost equal in terms of RMSE on the inner validation subset (25% of the total dataset).

When predicting deadwood biomass on the training dataset without an inner validation subset, there was a lower RMSE – 1.3 t ha^{-1} , or 16% of the mean deadwood biomass stock. According to the validation results, all 10 k -NN models have shown rather similar performance: RMSE ranges from 1.42 to 1.51 t ha^{-1} . The lowest RMSE was achieved by the model with $k = 1$ (1.42 t ha^{-1} , or 17% of the mean deadwood biomass stock). The best GBM and k -NN

models thus were used to impute on the spatial dataset (Figure 7).

Importantly, low deadwood biomass values ($2\text{--}3\text{ t ha}^{-1}$) were predicted by the GBM model (Figure 7b) as two or three times higher. Therefore, the k -NN model

(Figure 7a) visually has produced more “outliers” throughout the graph. A comparison of mean (with standard deviations, SD) and total deadwood biomass outputs for the entire study area is given in Table 4.

Table 4. Comparison of mean and total values of deadwood biomass stock within the study area produced by the best performing models.

Type of model	Tree species				
	ALGL	BEPE	OTHER	PISY	POTR
Mean $\pm$ SD , t ha^{-1}					
k -NN	7.0 ± 3.4	8.5 ± 1.9	8.1 ± 2.8	8.6 ± 1.9	8.1 ± 2.1
GBM	6.8 ± 2.3	8.0 ± 1.7	8.1 ± 1.7	8.4 ± 1.5	8.1 ± 2.1
Total, thousands t					
k -NN	39.0	34.8	19.9	237.6	4.9
GBM	37.9	32.8	19.9	232.0	4.9

The total deadwood biomass stock within the study area produced by the k -NN model is 0.34 Mt, while the best GBM resulted in similar stock (0.33 Mt, Table 4). The highest differences in the mean stock between model outputs were found for silver

birch, and, with less extent, for Scots pine and black alder.

Maps of predicted dead biomass stock for test landscapes (with 95% CI) are given in Figure 8, with a respective comparison in Table 5.

Figure 8. Test landscape with predicted deadwood biomass stock: imputed by the k -NN model (a) or estimated by the GBM model (b).

Table 5. Comparison of mean and total values of deadwood biomass stock for the test landscape at polygon level.

Type of model	Mean $\pm$ CI, t·ha ⁻¹	Total stock, t	Mean difference with reference per polygon, t·ha ⁻¹
Reference	7.8 $\pm$ 0.3	8011	–
k-NN	9.0 $\pm$ 0.1	9107	+4.5
GBM	8.7 $\pm$ 0.1	8996	+4.1

While there were no likely large differences between model outputs for the test landscape (Figure 8), there was at least 13% total overestimation of deadwood biomass stock compared to reference FPAM data (Table 5). However, mean differences at polygon (stand) level could reach half of the mean dead biomass stock.

Discussion

Challenges towards robust classification

To date, Ukraine is facing increasing challenges due to the lack of precise, statistically-based NFI data (Lesiv *et al.*, 2018). A pilot project run in two regions (Sumy and Ivano-Frankivsk) a decade ago showed both new evidence (forest biometric parameters of stands located outside state-owned forests, deadwood stocks) on the state of local Ukrainian forests and a strong demand for further expanding such inventory (Myroniuk, 2018). However, Ukraine does not have a full national forest inventory due to a number of reasons (Bilous *et al.*, 2019). Since the quality of ancillary data strictly defines the performance of any classification of remote sensing data, uncertainties and mistakes always occur while working with imprecise data. In this study, FPAM data also had a limited temporal frame (2010–2011), so only Landsat-5 TM imagery could be used for land cover classification and further processing, while the Landsat-8 Operational Land Imager archive has been available for years. FPAM data in Ukraine is also commonly recognized as biased and limited in precision, strictly focusing on management purposes

and neglecting the ecological functions of forests (Lesiv *et al.*, 2018). One possible suggestion is to carry out independent research of forest inventory (Bilous *et al.*, 2017). However, the use of such an approach for larger areas as proposed in this study requires substantial time and funds.

The spatial resolution of Landsat-5 TM imagery can still be considered as sufficient for the classification models while using the tools for validation purposes like Google Earth and Collect Earth. The cloud-masked mosaic, even with a limited period (one vegetation season in 2011), reliably produces a multispectral dataset without noises. The forest RF-based mask produced in this study may be compared with the forest cover map proposed by Global Forest Change (GFC, Hansen *et al.*, 2013). The GFC mask was utilized with a 40% canopy closure threshold, which is recognized as a suitable value for distinguishing forests in Ukraine (Myroniuk, 2018). The forested area according to this data was 31.2%, thus closely matching the value produced by our RF model (32.0%).

Some challenges linked to local geographical and biophysical features were met in this study. Linear and relatively narrow forest stands are common in Ukraine (shelterbelts, across roads and rivers, etc.). Myroniuk (2018) reported that the spatial resolution of Landsat bands is too coarse for proper classification of such forests surrounded by fields (Figure 9b). However, utilization of the NDWI as a predictor can be sufficient to provide reliable distinction of forest patches in river banks from water bodies (Figure 9a) and thus create a more precise forest mask. This task can be solved

Figure 9. Subsets of predicted forest cover within the study area vs. actual linear stands across water bodies (a–b) and pathways (c–d). The green mask is produced by the RF model and the purple mask is created using GFC data.

by using a dense Landsat time series (Cohen & Goward, 2004).

Because of the sparse distribution of silver birch stands within the study area, tree species classification (Table 3) has shown strong biases in distinction of this species from Scots pine forests. Lakyda *et al.* (2019) indicated that there was uncertainty in distinguishing between black alder and silver birch, since these species had similar spectral reflectance. This issue was solved by adding soil maps as an ancillary dataset, while in this study the non-spectral predictor DEM was important for delineating stands of black alder.

Reliability of deadwood estimation

Successful use of the k -NN method for forest inventory purposes throughout the world (McRoberts, 2012) has proven its capability to provide robust data imputation. However, this approach is sensitive to the

quality of field-sampled data (Tomppo *et al.*, 2017). Gradient boosting machines are commonly used for classification purposes and often compared to another ensemble method – RF, so there is some lack of information for tests of the imputation performance (Dube *et al.*, 2014). In this study we have observed that GBM can slightly outperform the k -NN model in terms of accuracy (Figure 7). On the other hand, k -NN does not require control of as many modelling parameters, so it is more user-friendly. The GBM tuning algorithm needs several hours to run depending on the number of hyperparameter combinations and computation power. However, it allows inner validation using a training fraction parameter, k -fold cross-validation and stochastic nature through the bagging fraction hyperparameter. Because of these advantages, XGBoost (Chen & Guestrin, 2016), the next extension of the AdaBoost

(and thus GBM) framework, has shown the best accurate results in various data science competitions working with structured datasets. Relatively low attention to GBM capabilities in forest imputation tasks requires further utilization of this method to obtain more robust woody stock and biomass predictions. Possible limitations to such use, compared to the simple and classic k -NN method, could be associated with local spectral features, the size and quality of the training dataset, nature of the satellite spatial and temporal resolution.

In this study we used linear allometric models to obtain data on deadwood biomass stocks. These models were developed based on the biophysical parameters of a growing stand and thus are strictly linked to natural mortality processes in different successional stages and reflect local site productivity. However, such models cannot predict deadwood stocks after abrupt events (natural disturbances at different scales) and different management considerations common in Ukraine (Bilous *et al.*, 2019). A network of sample plots with a reasonable allocation scheme can be harnessed for this issue but requires data from statistical national inventory (currently available only for two regions in Ukraine), global datasets, such as Schepaschenko *et al.* (2019), or substantial costs to carry out the LiDAR measurements (Gonzalez *et al.*, 2018). While we estimated deadwood indirectly through satellite images, more precise data could be directly obtained with the use of laser scanning (Zald *et al.*, 2016).

Conclusions

Deadwood stocks are strictly related to various issues ranging from biodiversity conservation to carbon management, but hitherto are not reflected in operational forest management data. However, those can be considered in the forest inventory agenda of countries with developing or transition economy utilizing free available

remote sensing data. To meet climate mitigation and ecosystem diversity maintenance targets, there is credible capability to apply a bundle of novel ensemble machine learning methods, e.g. Random Forest and Gradient Boosting Machines, with satellite multispectral data. These non-parametric approaches, together with the traditional k -NN regression estimator, could be utilized in different cases, varying in their interpretability, tuning suitability, and biological linkages with real data.

Assessment and mapping tasks of deadwood in Ukraine require urgent implementation of a national statistical inventory, preferably with utilization of diverse sources of data (like point clouds provided by LiDAR). Also, the proposed approach matches with the necessity to estimate dead biomass, while this ecosystem compartment is crucial to maintain forest resilience under the changing global climate.

Acknowledgements. This research was partially funded by the Ministry of Education and Science of Ukraine (Grant Nos. 110/5-pr-2018, Kyiv, Ukraine). Authors acknowledge Elijah Kelsey for his language support.

References

- Angelstam, P., Kuuluvainen, T. 2004. Boreal forest disturbance regimes, successional dynamics and landscape structures – a European perspective. – *Ecological Bulletins*, 51, 117–136.
- Belgiu, M., Drăguț, L. 2016. Random forest in remote sensing: a review of applications and future directions. – *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31.
- Bilous, A., Matsala, M., Radchenko, V., Matiashuk, R., Boyko, S., Bilous, S. 2019. Coarse woody debris in mature oak stands of Ukraine: carbon stock and decomposition features. – *Forestry Ideas*, 1(57), 196–219.
- Bilous, A., Myroniuk, V., Holiaka, D., Bilous, S., See, L., Schepaschenko, D. 2017. Mapping growing stock volume and forest live biomass: a case study of the Polissya region of Ukraine. – *Environmental Research Letters*, 12, e105001.
- Breiman, L. 2001. Random Forests. – *Machine Learning*, 45(1), 5–32.

- Chen, T., Guestrin, C. 2016. XGBoost: a scalable tree boosting system. – KDD' 16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, Aug. 2016. San Francisco, California, 785–794.
- Civco, D.L. 1993. Artificial neural networks for land-cover classification and mapping. – International Journal of Geographical Information Systems, 7(2), 173–186. <https://doi.org/10.1080/02693799308901949>.
- Cohen, W.B., Goward, S.N. 2004. Landsat's role in ecological applications of remote sensing. – BioScience, 54(6), 535–545. [https://doi.org/10.1641/0006-3568\(2004\)054\[0535:LRIEA O\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2004)054[0535:LRIEA O]2.0.CO;2).
- Cohen, W.B., Yang, Z., Kennedy, R. 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync – Tools for calibration and validation. – Remote Sensing of Environment, 114(12), 2911–2924.
- Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. – Remote Sensing of Environment, 37(1), 35–46.
- Crookston, N.L., Finley, A.O. 2008. yalmpute: An R package for k-NN imputation. – Journal of Statistical Software, 23(10). 16pp.
- Dube, T., Mutanga, O., Elhadi, A., Ismail, R. 2014. Intra-and-inter species biomass prediction in a plantation forest: testing the utility of high spatial resolution spaceborne multispectral RapidEye sensor and advance machine learning algorithms. – Sensors, 14(8), 15348–15370. <https://doi.org/10.3390/s140815348>.
- Evans, J.S., Cushman, S.A. 2009. Gradient modelling of conifer species using random forests. – Landscape Ecology, 24, 673–683.
- Franklin, J.F., Spies, T.A., Pelt, R.V., Carey, A.B., Thornburgh, D.A., Berg, D.R., Lindenmayer, D.B., Harmon, M.E., Keeton, W.S., Shaw, D.C., Bible, K., Chen, J. 2002. Disturbances and structural development of natural forest ecosystems with silvicultural implications, using Douglas-fir forests as an example. – Forest Ecology and Management, 155, 399–423. [https://doi.org/10.1016/S0378-1127\(01\)00575-8](https://doi.org/10.1016/S0378-1127(01)00575-8).
- Fu, Y., He, H.S., Hawbaker, T.J., Henne, P.D., Zhu, Z., Larsen, D.R. 2019. Evaluating k-Nearest Neighbor (kNN) imputation models for species-level aboveground forest biomass mapping in Northeast China. – Remote Sensing, 11(17). <https://doi.org/10.3390/rs11172005>.
- Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R. 2006. Random Forests for land cover classification. – Pattern Recognition Letters, 27(4), 294–300.
- Gonzalez, R.S., Latifi, H., Weinacker, H., Dees, M., Koch, B., Heurich, M. 2018. Integrating LiDAR and high-resolution imagery for object-based mapping of forest habitats in a heterogeneous temperate forest landscape. – International Journal of Remote Sensing, 39(23), 8859–8884. <https://doi.org/10.1080/01431161.2018.1500071>.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. – Remote Sensing of Environment, 202, 18–27.
- Greenwell, B., Boehmke, B., Cunningham, J. 2019. GBM: generalized boosted regression models. CRAN repository, version 2.1.5. [WWW document]. – URL <https://CRAN.R-project.org/package=gbm>. [Accessed 20 February 2020].
- Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G. 2013. High-resolution global maps of 21-century forest cover change. – Science, 342, 850–853. <https://doi.org/10.1126/science.1244693>.
- Hayes, M.M., Miller, S.N., Murphy, M.A. 2014. High-resolution landcover classification using Random Forest. – Remote Sensing Letters, 5(2), 112–121. <https://doi.org/10.1080/2150704X.2014.882526>.
- Jhonnerie, R., Siregar, V.P., Nababan, B., Prasetyo, L.B., Wouthuyzen, S. 2015. Random forest classification for mangrove land cover mapping using Landsat 5 TM and ALOS PALSAR imageries. – Procedia Environmental Sciences, 24, 215–221. <https://doi.org/10.1016/j.proenv.2015.03.028>.
- Kauth, R.J., Thomas, G.S. 1976. The tasseled cap – a graphic description of the spectral-temporal development of agricultural crops as seen by LANDSAT. – Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, USA, June/July 1976. Indiana, 4B–41 to 4B–51.
- Kennedy, R.E., Yang, Z., Cohen, W.B. 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr – Temporal segmentation algorithms. – Remote Sensing of Environment, 114(12), 2897–2910.
- Kuhn, M. 2008. Building predictive models in R using the caret package. – Journal of Statistical Software, 28(5). <https://doi.org/10.18637/jss.v028.i05>.
- Lakyda, P., Shvidenko, A., Bilous, A., Myroniuk, V., Matsala, M., Zibtsev, S., Schepaschenko, D., Holiaka, D., Vasylyshyn, R., Lakyda, I., Diachuk, P., Kraxner, F. 2019. Impact of disturbances on the carbon cycle of forest ecosystems in Ukrainian Polissya. – Forests, 10, 337. <https://doi.org/10.3390/f10040337>.

- Latifi, H., Fassnacht, F.E., Hartig, F., Berger, C., Hernandez, F., Corvalan, P., Koch, B. 2015. Stratified aboveground forest biomass estimation by remote sensing data. – *International Journal of Applied Earth Observation and Geoinformation*, 38, 229–241.
- Lesiv, M., Shvidenko, A., Schepaschenko, D., See, L., Fritz, S. 2018. A spatial assessment of the forest carbon budget for Ukraine. – *Mitigation and Adaption Strategies for Global Change*, 24(6), 985–1006.
- Li, M., Im, J., Beier, C. 2013. Machine learning approaches for forest classification and change analysis using multi-temporal Landsat TM images over Huntington Wildlife Forest. – *GIScience & Remote Sensing*, 50(4), 361–384. <https://doi.org/10.1080/15481603.2013.819161>.
- Liaw, A. Wiener, M. 2002. Classification and regression by randomForest. – *R News*, 2(3), 18–22.
- Lim, K., Treitz, P., Wulder, M., St-Onge B., Flood M. 2003. LiDAR remote sensing of forest structure. – *Progress in Physical Geography: Earth and Environment*, 27(1), 88–106. <https://doi.org/10.1191/0309133303pp360ra>.
- Liu, W., Song, C., Schroeder, T.A., Cohen, W.B. 2008. Predicting forest successional stages using multitemporal Landsat imagery with forest inventory and analysis data. – *International Journal of Remote Sensing*, 29(13), 3855–3872.
- Maxwell, A.E., Warner, T.A., Fang, F. 2018. Implementation of machine-learning classification in remote sensing: an applied review. – *International Journal of Remote Sensing*, 39(9), 2784–2817. <https://doi.org/10.1080/01431161.2018.1433343>.
- McRoberts, R.E. 2012. Estimating forest attribute parameters for small areas using nearest neighbors techniques. – *Forest Ecology and Management*, 272, 3–12. <https://doi.org/10.1016/j.foreco.2011.06.039>.
- Myroniuk, V. 2018. Forest cover mapping using Landsat-based seasonal composite mosaics. (Класифікація лісового покриття за сезонними композитними мозаїками Landsat). – *Scientific Bulletin of UNFU*, 28(1), 28–33. (In Ukrainian).
- Lang, M., Arumäe, T., Lück, 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. <https://doi.org/10.2478/fsmu-2014-0010>.
- 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. <https://doi.org/10.2478/fsmu-2018-0001>.
- Nitze, I., Barrett, B., Cawkwell, F. 2015. Temporal optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series. – *International Journal of Applied Earth Observation and Geoinformation*, 34, 136–146. <https://doi.org/10.1016/j.jag.2014.08.001>.
- Oeser, J., Pflugmacher, D., Senf, C., Heurich, M., Hostert, P. 2017. Using intra-annual landsat time series for attributing forest disturbance agents in Central Europe. – *Forests*, 8, 251. <https://doi.org/10.3390/f8070251>.
- Pflugmacher, D., Cohen, W.B., Kennedy, R.E. 2012. Using Landsat-derived disturbance history (1972–2010) to predict current forest structure. – *Remote Sensing of Environment*, 122, 146–165. <https://doi.org/10.1016/j.rse.2011.09.025>.
- R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [WWW document]. – URL <http://www.R-project.org/>. [Accessed 20 February 2020].
- Ruefenacht, B., Finco, M.V., Nelson, M.D., Czaplewski, R., Helmer, E.H., Blackard, J.A., Holden, G.R., Lister, A.J., Salajano, D., Weyerermann, D., Winterberger, K. 2008. Conterminous U.S. and Alaska forest type mapping using forest inventory and analysis data. – *Photogrammetric Engineering & Remote Sensing*, 11, 1379–1388. <https://doi.org/10.14358/PERS.74.11.1379>.
- Sachdeva, S., Bhatia, T., Verma, A.K. 2018. GIS-based evolutionary optimized Gradient Boosted Decision Trees for forest fire susceptibility mapping. – *Natural Hazards*, 92, 1399–1418. <https://doi.org/10.1007/s11069-018-3256-5>.
- Schepaschenko, D., Chave, J., Phillips, O.L., Lewis, S.L., Davies, S.J., Réjou-Méchain, M., Sist, P., Scipal, K., Perger, C., Herault, B., Labrière, N., Hofhansl, F., Affum-Baffoe, K., Aleinikov, A., Alonso, A., Amani, C., Araujo-Murakami, A., Armston, J., Arroyo, L., Ascarrunz, N., Azevedo, C., Baker, T., Balazy, R., Bedeau, C., Berry, N., Bilous, A.M., Bilous, S.Y., Bissiengou, P., Blanc, L., Bobkova, K.S., Braslavskaya, T., Brienen, R., Burslem, D.F.R.P., Condit, R., Cuni-Sanchez, A., Danilina, D., del Castillo Torres, D., Derroire, G., Descroix, L., Sotta, E.D., d'Oliveira, M.V.N., Dresel, C., Erwin, T., Evdokimenko, M.D., Falck, J., Feldpausch, T.R., Foli, E.G., Foster, R., Fritz, S., Garcia-Abril, A.D., Gornov, A., Gornova, M., Gothard-Bassébé, E., Gourlet-Fleury, S., Guedes, M., Hamer, K.C., Susanty, F.H., Higuchi, N., Coronado, E.N.H., Hubau, W., Hubbell, S., Ilstedt, U., Ivanov, V.V., Kanashiro, M., Karlsson, A., Karminov, V.N., Killeen, T., Koffi, J.-C.K., Konovalova, M., Kraxner, F., Krejza, J., Krisnawati, H., Krivobokov, L.V., Kuznetsov, M.A., Lakyda, I., Lakyda, P.I., Licona, J.C., Lucas, R.M., Lukina, N., Lussetti, D., Malhi, Y., Manzanera, J.A., Marimon, B., Junior, B.H.M., Martinez, R.V., Martynenko, O.V., Matsala,

- M., Matyashuk, R.K., Mazzei, L., Memiaghe, H., Mendoza, C., Mendoza, A.M., Moroziuk, O.V., Mukhortova, L., Musa, S., Nazimova, D.I., Okuda, T., Oliveira, L.C., Ontikov, P.V., Osipov, A.F., Pietsch, S., Playfair, M., Poulsen, J., Radchenko, V.G., Rodney, K., Rozak, A.H., Ruschel, A., Rutishauser, E., See, L., Shchepashchenko, M., Shevchenko, N., Shvidenko, A., Silveira, M., Singh, J., Sonké, B., Souza, C., Stereńczak, K., Stonozhenko, L., Sullivan, M.J.P., Szatniewska, J., Taedoumg, H., ter Steege, H., Tikhonova, E., Toledo, M., Trefilova, O.V., Valbuena, R., Gamarra, L.V., Vasiliev, S., Vedrova, E.F., Verhovets, S.V., Vidal, E., Vladimirova, N.A., Vleminckx, J., Vos, V.A., Vozmitel, F.K., Wanek, W., West, T.A.P., Woell, H., Woods, J.T., Wortel, V., Yamada, T., Nur Hajar, Z.S., Zo-Bi, I.C. 2019. The Forest Observation System, building a global reference dataset for remote sensing of forest biomass. – *Scientific Data*, 6(1), 198. <https://doi.org/10.1038/s41597-019-0196-1>.
- Seidl, R., Rammer, W., Spies, T.A. 2014. Disturbance legacies increase the resilience of forest ecosystem structure, composition, and functioning. – *Ecological Applications*, 24, 2063–2077. <https://doi.org/10.1890/14-0255.1>.
- Shvidenko, A., Buksha, I., Krakovska, S., Lakyda, P. 2017. Vulnerability of Ukrainian forests to climate change. – *Sustainability*, 9, 1152. <https://doi.org/10.3390/su9071152>.
- Thanh Noi, P., Kappas, M. 2018. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector machine classifiers for land cover classification using Sentinel-2 imagery. – *Sensors*, 18(1), 18. <https://doi.org/10.3390/s18010018>.
- Tomppo, E., Kuusinen, N., Mäkisara, K., Katila, M., McRoberts, R.E. 2017. Effects of field plots configuration on the uncertainties of ALS-assisted forest resource estimates. – *Scandinavian Journal of Forest Research*, 32(6), 488–500. <https://doi.org/10.1080/02827581.2016.1259425>.
- Zald, H.S.J, Wulder, M.A., White, J.C., Hilker, T., Hermosilla, T., Hobart, G.W., 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 Sensing of Environment*, 176, 188–201. <https://doi.org/10.1016/j.rse.2016.01.015>.
- Zimble, D.A, Evans, D.L., Carlson, G.C., Parker, R.C., Grado, S.C, Gerard, P.D. 2003. Characterizing vertical forest structure using small-footprint airborne LiDAR. – *Remote Sensing of Environment*, 87, 171–182. [https://doi.org/10.1016/S0034-4257\(03\)00139-1](https://doi.org/10.1016/S0034-4257(03)00139-1).

Received September 16, 2020, revised November 14, 2020, accepted November 16, 2020