Forest Stand Volume Estimation by Species from Sentinel-2 and LiDAR Data Using Regression Models

Juris Sinica-Sinavskis
Institute of Electronics and Computer Science
Riga, Latvia
jss@edi.lv

Gunta Grube
Institute of Electronics and Computer Science
Riga, Latvia
gunta.grube@edi.lv

Abstract—**The paper presents a novel forest stand volume (FSV) estimation approach based on remote sensing (RS) data when the forest inventory data used as reference are limited. The proposed approach consists of several steps, such as filtering of existing inventory data, identifying individual tree tops from the Canopy Height Model (CHM), classifying dominant tree species from Sentinel-2 data, and creating the polynomial regression model for stand volume estimation based on training data. The study area was located in the Zemgale region of Southeast Latvia, where the dominant tree species are Scots pine (*Pinus sylvestris* L.), Norway spruce (*Picea abies* (L.) Karst.), birch (*Betula pendula* Roth, *Betula pubescens* Ehrh.) and black alder (*Alnus glutinosa* (L.) Gaertn.). The FSV (m3/ha) for each dominant species was estimated, and the accuracy against the harvester data was evaluated by calculating the root mean square error (*RMSE*). Additionally, a cross-validation was performed using sparse and partially imprecise inventory data, and the *RMSE* errors were less than 20% for pine, 22% for spruce, 28% for birch, and 23% for black alder. In general, the developed approach can be used with species for which there is a sufficient number of inventory compartments in the analysis region where these species dominate. The proposed approach can be used in automatic workflows estimating forest inventory parameters from RS data.**

Keywords— Forest stand volume estimation, Canopy Height Model, Multispectral imaging, Remote sensing.

I. Introduction

FSV assessment is one of the key tasks of forest inventory performed in Latvia at least once in 20 years [1]. In many European countries including Latvia, forest inventory is performed by a taxator walking around the forest stands and estimating the main taxation indicators manually. Due to high labor costs, estimates of inventory indicators are obtained according to some general methodologies, without measuring each tree. On the contrary, several European countries are trying to introduce new RS-based solutions for obtaining inventory indicators. For example, Finland is coping very well with that, probably due to the domination of a fairly homogeneous pine and spruce forest [2-4]. In Ukraine, on the other hand, RS is used to divide forests into homogeneous units (primary units of forest inventory). Then the forest parameters (i.e. tree species composition, age, average height and diameter, growing stock volume, etc.) are assigned by trained staff on the ground through visual estimation or measurements [5]. In the Baltic States, such a solution has not yet been implemented, as it requires sophisticated data processing due to the extensive and diverse training data, as well as very high legal requirements for the accuracy of inventory parameter estimates (± 20% maximum error) [1].

The use of RS is hampered by the high required precision and the diversity of tree species in hemiboreal mixed forests [6]. Such forest inventories with direct use of freely available imagery (i.e., Landsat 8 and Sentinel-2) might not be compliant due to the moderate resolution products (i.e., 10-30m) and limited correlation of spectral variables with forest structural properties [7]. Combining spectral variables and airborne laser scanning (ALS) [6] or unmanned aerial vehicles (UAV) data [7-8] might lead to an increase of the estimation precision of the key forest parameters. ALS data are often used to obtain a tree height model for further segmentation or, in combination with statistical methods, to estimate inventory parameters [2, 10-11]. For example, Noordermeer et al. developed a specific fitted regression predictive model based on tree height for spruce and pine stands using repeated ALS data [12]. Several authors [3-5, 11] have used different algorithms for estimating inventory parameters together with field data. Using data from UAV, e.g., Mosaicmill company offers forest inventory without fieldwork [8]. It could be an effective way to perform forest inventory, compared to fieldwork [9].

Depending on the availability of RS data, appropriate forest inventory estimation models can be developed. The RS data specifics should be carefully considered to obtain unbiased estimates of forest parameters at the level of accuracy desired by the user. Satellite images are still the most extensively used RS technology due to the low cost and frequent returns. Nevertheless, airborne laser scanning (ALS) combined with spectral and photogrammetric data is known to be particularly useful for the assessment of forest structural attributes. Research activities carried out in the past years [13], demonstrated estimated growing stock volume with a total *rRMSE* of 13.4% and species-specific volumes predicted with *rRMSE* of 36.6%, 46.5%, and 84.9% for spruce, pine, and deciduous species, respectively.

Another approach such as local weighted regression models is used for FSV estimation and assumes a general relationship between stock and tree crown height [14].

In this paper, we offer an original stock assessment solution consisting of tree species classification and tree identification procedures and using a polynomial regression model with forest inventory data as support data.

978-1-6654-6295-2/22/$31.00 ©2022 European Union

Furthermore, the calculated forest stock values were validated against the harvester data.

II. USED DATA

LiDAR data were obtained from the Latvian Geospatial Information Agency. The acquired point cloud contains a network of points with coordinates (X, Y, H) in the LKS-92 coordinate system (EPSG:3059). The data for the western part of the study site were obtained in 2015, while for the eastern part in 2017.

The Level-1C Sentinel-2 images of the study area in Latvia (see Fig.1) used in this research were downloaded from the Copernicus Open Access Hub [15]. They were acquired by the Sentinel-2A satellite on 25 April 2019 (non-leaf period) and 4 June 2019, and the Sentinel-2B satellite on 27 September 2019. Ten Sentinel-2 bands were used, featuring 10- and 20-m spatial resolution, and obtained in the visible, near-infrared, and shortwave infrared spectral ranges (bands B2, B3, B4, B5, B6, B7, B8, B8a, B11, and B12). All the bands were resampled to the 10-m resolution. The images were combined to process them together as a 30-band image.

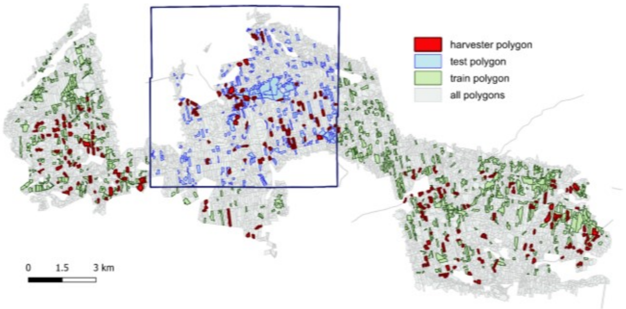


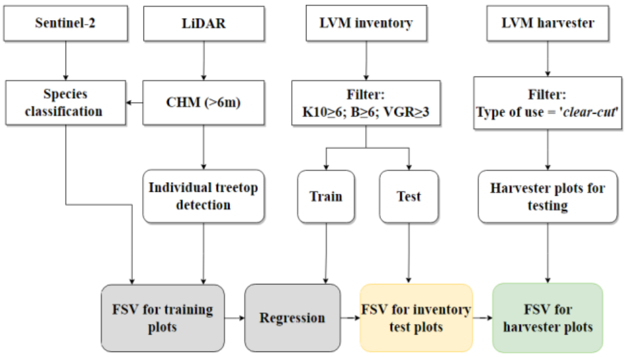
Fig. 1. Study area located in the Zemgale region of Southeast Latvia (center coordinated 56.50°N, 25.00°E)

We exploited the Regular stand-wise forest **inventory** (RFI) [1] **data** (see Fig.1) obtained from the JSC "Latvia's State Forests" (LVM), and filtered clear-cut forest stands for the study area by using a clearcutting mask [16], 9599 in total. The RFI data contained coordinates of the forest stands (inventory polygons), tree species composition data, site fertility class, forest age group (VGR) in range (1..5), the total wood volume in cubic meters per hectare (VNOG), share coefficient (K10) for the dominant species in range (0..10) indicating the proportion of the number of trees in a forest stand, first storey forest density relative to the 'normal' density (B10) in range (0..10). RFI data filtering by K10 ≥ 8; B10 ≥ 6; VGR ≥ 3 resulted in 1382 selected plots which were split into 971 (70%) training and 411 (30%) test data sets for internal use, see Fig.1. Four tree species dominating in the analysis area were selected: Scots pine (*Pinus sylvestris* L.), Norway spruce (*Picea abies* (L.) Karst.), birch (*Betula pendula* Roth and *Betula pubescens* Ehrh.), and black alder (*Alnus glutinosa* (L.) Gaertn.).

Harvester data were available for individual trees in forest compartments within the study site carved during the winter of 2020/2021. Tree species, felling location, use type, and tree volume (m3) were recorded as variables. We used only harvester data from clear-cut forest compartments. Then we related the data to the inventory polygons obtaining the total harvested stock volume (m3/ha) ($V_{\mathrm{Harv\_Total}}$) by species. The number of clear-cut forest compartments with dominating birch and black alder species was relatively low, hindering a statistically significant assessment. In total, the results were validated by 278 plots with the harvester data provided by LVM.

III. METHOD

The main workflow of the stock volume estimation contains three separate procedures: identification of tree canopy peaks, tree species classification, and preparation and exploitation of a regression model for each species, see below. The approach proposed is based on the assumption that some sparse and partially imprecise forest inventory data are available. The sparse and outdated forest inventory data contain information about forest stands with attributes such as tree species, volume, height, etc. It is necessary to obtain full and updated information about the area, in particular, the FSV. The data processing workflow consists of several steps and is presented in Fig.2.



Sentinel-2
LiDAR
LVM inventory
LVM harvester
Species classification
CHM (>6m)
Filter: K10≥6; B≥6; VGR≥3
Filter: Type of use = 'clear-cut'
Individual treetop detection
Train
Test
Harvester plots for testing
FSV for training plots
Regression
FSV for inventory test plots
FSV for harvester plots

Fig. 2. General workflow of the forest stock volume (m3/ha) estimation method.

An image of size 1554x2725 (belonging to the study area shown in Fig.1) processing took several hours, the most computational resources are used in the step of species classification which exploited Dynland clustering [17].

A. Identification of tree canopy peaks

To identify the tree canopy peaks, we have used the CanopyMaxima algorithm from Fusion/LDV software [18]. The following algorithms were used for CanopyMaxima tree detection from LiDAR data: GroundFilter, GridSurfaceCreate; CanopyModel, CanopyMaxima, and LDA2ASCII, DTM2ASCII for data format conversion. CanopyMaxima uses a canopy height model to identify local maxima using a variable-size evaluation window. The window size (WS) is based on the canopy height. This tool can only identify dominant and codominant trees in the upper canopy for some forest types. If the WS is too small, a lot of commission errors will occur. If too large, many treetops will be omitted [18].

The LiDAR point cloud was filtered to identify returns from the ground; then a gridded surface model was created with a spatial resolution of 1 m as it was the most suitable resolution for the available data; then a CHM with the spatial resolution of 0.5 m was created and filtered using the median 5x5 pixels filter as this option produced the highest accuracy regarding the trees' detection. We chose the individual tree detection algorithm of CanopyMaxima to identify trees above *h*=6 m and we set the window size *s* according to the formula (1):

$$s=0.8+0.05 \cdot h \tag{1}$$

TABLE I. The FSV results of the *R*, *RMSE*, and *rRMSE* that were computed by the PR model

Attributes	Polygon count	*R*	*RMSE*	*rRMSE*, %
FSV_{Pine}	80	0.78	52	14
FSV_{Spruce}	146	0.76	59	15
FSV_{Birch}	30	0.52	71	19
$FSV_{BlackAlder}$	22	0.59	77	22

B. Tree species classification

To perform classification, we clustered the Sentinel-2 images using the Dynland clustering algorithm [17]. Tree species classes were automatically assigned to the obtained clusters by using the algorithm described in [19]. To perform clustering of the whole study area, each second pixel was clustered on both axes using the Dynland algorithm [17], and other pixels were put into the formed 265 clusters on a spectral similarity basis using k-nearest neighbors search.

C. Preparing and exploiting regression model

We used 963 inventory plots to prepare a regression model. Before performing the polynomial regression (PR), the basal area (m2) and forest stock volume (m3/ha) of each inventory plot was computed as defined in [20].

The general formulas for calculating the basal area and forest stock volume are as follows:

$$G = 0.7854 \cdot \left(\frac{H}{100}\right)^2 \cdot N \tag{2}$$

where *G* - basal area (m2); *H* - average tree height (m) used instead of diameter breast height (DBH) due to strong relation between mean DBH and tree height (with bias of 13% for Scots pine, 11% for Norway spruce, 7% for Silver birch, and 3% for Black alder) [21], and *N* - number of trees.

$$V = \frac{k \cdot G \cdot (H+4)}{A} \tag{3}$$

where *V* - forest stock volume (m3/ha); *k* is species-specific coefficient (pine: 0.390, spruce: 0.415, birch: 0.385, black alder: 0.400) [20]; *G* - basal area (m2); *A* - the area of plot (ha).

Basal area and stock volume were calculated for each species separately. We used individual trees and height from LiDAR and classification from Sentinel-2 to define species distribution and its area in each inventory plot.

We used a 2nd-degree polynomial relationship between the response and predictor variables to avoid complexity and overfitting. The PR model was found to be suitable for the estimation of FSV at the study site, based on trial-and-error using correlation measures. This model fits a connection between the dependent and independent variables as a 2nd-degree polynomial using the method of polynomial least squares [22]:

$$\begin{aligned} \mathrm{E}(Y) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \\ &+ \beta_4 X_1 X_2 + \beta_5 X_2^2 + \varepsilon \end{aligned} \tag{4}$$

where *Y* is the dependent variable, E*(Y)* is the expected value of *Y*, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_5$ are the regression coefficients of predictors X_1, X_2, and ε is the residual error. We used the inventory first storey stock volume (m3/ha) as our response *Y*. We computed first storey stock volume (m3/ha) and first storey average height obtained from individual trees and LiDAR CHM as our predictors X_1 and X_2.

IV. RESULTS AND ACCURACY ASSESSMENT

Performance of the models was examined using the correlation coefficient (*R*) (5), root-mean-square-error(*RMSE*) (6), and relative root-mean-square-error % (*rRMSE*) (7) on the training and harvester validation plots:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2}}, \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{6}$$

$$rRMSE = RMSE / \bar{y} \cdot 100, \tag{7}$$

where x_i represents the observed values for plot *i*; $\bar{x}$ - the average observed values for all plots; y_i - the estimated values for plot *i*; $\bar{y}$ - the average estimated values for all plots; *n* is the number of plots; and *i* is the sample number.

The model was cross-validated to compare the results obtained from different groups of predictor variables. The *10-fold* cross-validation (CV) involves splitting 1382 selected plots into 10 subsets. One subset from each fold was used to test the model's performance, while all other nine subsets were used for training the model. The results indicated that the *rRMSE* values for the PR model range from 10 to 28% (see Fig. 3) with an average *RMSE* score of 50, 57, 57, and 50 m3/ha for pine, spruce, birch, and black alder polygons, respectively.

Results were the most accurate for estimating pine (*R*= 0.78, *RMSE*=52 m3/ha) and spruce (*R*=0.76; *RMSE*=59 m3/ha) FSV, followed by birch (*R*=0.52; *RMSE*=71 m3/ha) and black alder (*R*=0.59; *RMSE*=77 m3/ha) polygons, see Tab. 1. Correlation graphs between total harvested volume $V_{\mathrm{Harv\_Total}}$ and total predicted FSV by dominant species are demonstrated in Fig.4. Fig. 5 shows predicted FSV maps in the range 0 – 450 m3/ha. In general, underestimation of FSV with respect to inventory data is observed; it is understandable as the individual trees are counted in CHM without the possibility to take into account the second storey of the forest.

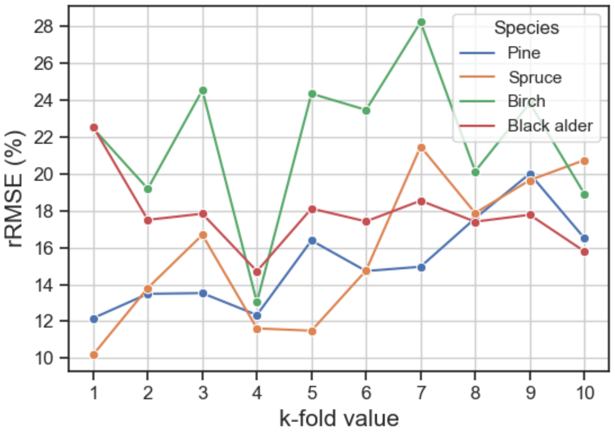


Fig. 3. Sensitivity of the *rRMSE* values to different 10-fold values fitted with the PR method.

V. Discussion and Further Work

The available solutions from remote sensing for the evaluation of forest inventory parameters in the scientific literature can be divided as follows: 1) solutions that use only laser data together with reference data; 2) solutions that use only optical data with reference data; 3) solutions that use both laser and optical data with reference data. In our opinion, the best solution is to use as much data as possible, taking into account the advantages and disadvantages of each data type to evaluate specific inventory parameters. Such solutions are summarized in a literature review article [9].

Most of the previous works, such as [2, 3], have presented methods for total forest stock volume estimation disregarding species distribution. On the contrary, this study demonstrates forest structural attribute assessment from Sentinel-2 and LiDAR data separately for four dominant species, using available inventory data as a reference. This is particularly beneficial for detailed forest inventory.

One of the problems of the proposed solution is the different acquisition times of LiDAR and Sentinel-2 data. It is obvious that the latest Sentinel-2 data should be used as the base for the method. Special procedures should be developed to diminish the impact of outdated LiDAR data. The proposed method can be supplemented with growth coefficients with respect to old LiDAR data [9, 20].

The accuracy of the proposed method was assessed against pure pine, spruce, birch, and black alder stands. A lower *rRMSE* was obtained for pine and spruce stands, 14% and 15%, respectively, followed by deciduous stands of birch (19%) and black alder (22%). It should be taken into account that the error rate in mixed forest stands can be significantly higher in mixed forests.

We believe that accuracy limitations can be reduced by using up-to-date LiDAR data with higher resolution or UAV-collected data. Sentinel-2 data from several seasons provide more information for the identification of tree species. However, the limited availability of cloud-free satellite images may impose restrictions on the use of the proposed FSV estimation approach.

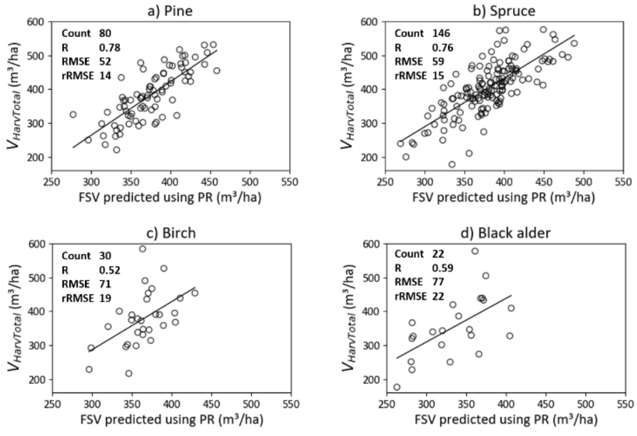


Fig. 4. Scatterplots of a total harvester (V_{Harv_Total}) vs total predicted FSV (m3/ha) using PR in test polygons, by dominant species.

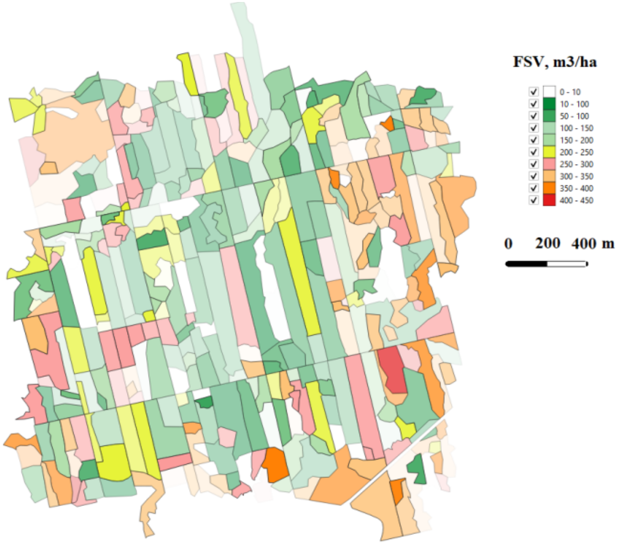


Fig.5. FSV (m3/ha) prediction maps derived from LiDAR CHM, Sentinel-2 multispectral images using sparse and outdated forest inventory data (center coordinates 56.50°N, 25.00°E)

Future work will include an examination of the potential of using UAV-collected data for forest inventory. The potential of classification and FSV estimation for additional species should also be explored.

VI. Conclusion

In this paper, we have proposed a novel stock volume estimation approach from Sentinel-2 multispectral images and LiDAR CHM using available sparse and outdated forest inventory data as the reference. Our approach is based on the recently proposed Dynland clustering algorithm, identification of tree canopy peaks, and using a regression model for stock volume estimation. Experiments showed that a lower *RMSE* can be achieved using polynomial regression on the chosen training set for forest stands. The highest correlations (*R*=0.78, *R*=0.76, *R*=0.52, *R*=0.59) for four tree species (pine, spruce, birch, and black alder) were obtained in this case. The proposed approach facilitates the estimation of other inventory parameters of forest stands separately for each species of interest.

Acknowledgment

The authors would like to thank the JSC Latvia's State Forests for providing the important data for this study, which is a part of an ERDF-funded project entitled "Satellite remote sensing-based forest stock estimation technology" (Project No. 1.1.1.1/18/A/165) and project "Remote sensing-based system for forest risk factor monitoring" (Project No. 1.1.1.1/19/A/128).

The authors would like to thank IECS Remsens group colleagues for their help in preparing the paper.

References

[1] Forest Inventory and State Forest Register Information Circulation Regulations, 2016 [Online]. Available: https://likumi.lv/ta/id/283091-meza-inventarizacijas-un-meza-valsts-registra-informacijas-aprites-noteikumi [Accessed 4-Jul-2022]

[2] J. Hyyppä, X. Yu, H. Hyyppä, M. Vastaranta, M. Holopainen, A. Kukko, H. Kaartinen, A. Jaakkola, M. Vaaja, J. Koskinen and P. Alho, "Advances in forest inventory using airborne laser scanning", *Rem. Sens.*, vol. 4, pp. 1190-1207, 2012.

[3] S. Wittke, X. Yu, M. Karjalainen, J. Hyyppa, E. Puttonen, "Comparison of two-dimensional multitemporal Sentinel-2 data

with three dimensional remote sensing data sources for forest inventory parameter estimation over a boreal forest", *Int. J. Appl. Earth Obs. Geoinformation*, vol. 76, pp. 167-178, 2019.

[4] P. Packalen, J. Strunk, T. Packalen, M. Maltamo, L. Mehtatalo, "Resolution dependence in an area-based approach to forest inventory with airborne laser scanning", *Rem. Sens. of Env.*, vol. 224, pp. 192-201, 2019.

[5] A. Bilous, V. Myroniuk, D. Holiaka, S. Bilous, L. See, "Mapping growing stock volume and forest live biomass: a case study of the Polissya region Ukraine, *Environmental Research Letters*, vol.12, no. 10, pp 1-13, 2017.

[6] M. Lang, L. Gulbe, A. Traskovs, A. Stepcenko, "Assessment of different estimation algorithms and remote sensing data sources for regional level wood volume mapping in hemiboreal mixed forests", *Baltic Forestry*, vol. 22, no. 2, pp. 283-296, 2016.

[7] S. Puliti, S. Saarela, T. Gobakken, G. Stahl, E. Næsset, "Combining UAV and Sentinel-2 auxiliary data for forest growing stock volume estimation through hierarchical model-based inference," *Rem. Sens. of Env.*, Vol. 204, pp. 485–497, 2018.

[8] Mosaicmill company "UAV inventory" [Online]. Available: https://www.mosaicmill.com/forestry/UAV_inventory.html [Accessed 4-Jul- 2022].

[9] P. Surov, K. Kuzelka, "Acquisition of Forest Attributes for Decision Support at the Forest Enterprise Level Using Remote-Sensing Techniques — A Review", Forests, vol. 10, no. 3, pp. 1-29, 2019.

[10] H. Latifi, F. E. Fassnacht, J. Müller, A. Tharani, S. Dech, M. Heurich, "Forest inventories by LiDAR data: A comparison of single tree segmentation and metric-based methods for inventories of a heterogeneous temperate forest", *Int. J. of App. Earth Obs. and Geoinf.*, vol. 42, pp. 162-174, October 2015.

[11] L. Gulbe, J. Zarins, I. Mednieks, "Automated Delineation of Microstands in Hemiboreal Mixed Forests Using Stereo GeoEye-1 Data", *Remote Sensing*, vol. *14, no.*6, 2022.

[12] L. Noordermeer, T. Gobakken, E. Nasset, O. M. Bollandsas, "Predicting and mapping site index in operational forest inventories using bitemporal airborne laser scanner data", *Forest Ecology and Management*, vol. 457, pp. 1-15, 2020.

[13] S. Puliti, T. Gobakken, H.O. Ørka and E. Næsset, "Assessing 3D point clouds from aerial photographs for species-specific forest inventories", *Scandinavian journal of forest research*, vol. 32, no. 1, pp. 68-79, 2017.

[14] F.Maselli, M.Chiesi, M.Mura, M.Marchetti, P.Corona, G.Chirici,"Combination of optical and LiDAR satellite imagery with forest inventory data to improve wall-to-wall assessment of growing stock in Italy", *Int. J. of App. Earth Observ. and Geoinf.*, vol. 26, pp. 377-386, February 2014.

[15] ESA. Copernicus Open Access Hub [Online]. Available: scihub.copernicus.eu [Accessed 4-Jul- 2022].

[16] G. Goldbergs, "Impact of Base-to-Height Ratio on Canopy Height Estimation Accuracy of Hemiboreal Forest Tree Species by Using Satellite and Airborne Stereo Imagery". *Remote Sensing*, vol. 13, no. 15, pp. 1-22, 2021.

[17] R. Dinuls and I. Mednieks, "Nonparametric classification of satellite images", in *Proc. of the 2018 International Conference on Mathematics and Statistics*, Porto, Portugal, July 15-17, pp. 64-68.

[18] R. McGaughey, "FUSION/LDV: Software for lidar data analysis and visualization," FUSION version 3.80, 2019.

[19] J. Sinica-Sinavskis, R. Dinuls, J. Zarins, I. Mednieks, "Automatic tree species classification from Sentinel-2 images using deficient inventory data", In *2020 17th Biennial Baltic Electronics Conference (BEC)*.

[20] I. Liepa. *Pieauguma maciba (Increment Science)*. Jelgava, LLU, 123 p, 1996.

[21] G. Prieditis, I. Smits, I. Arhipova, S. Daais, D. Dubrovskis, "Allometric Models for Predicting Tree Diameter at Breast Height," *Energy Syst. Sustain*, 4, pp.105-110, 2012.

[22] D. Bera, N. D. Chatterjee, S. Bera, "Comparative performance of linear regression, polynomial regression and generalized additive model for canopy cover estimation in the dry deciduous forest of West Bengal", *Rem. Sens. Applic.: Soc. and Env.*, Vol. *22*, No. 100502, 2021.