Ecological Informatics 50 (2019) 24–32

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam

An Thi Ngoc Dang[a], Subrata Nandy[b,*], Ritika Srinet[b], Nguyen Viet Luong[c], Surajit Ghosh[d], A. Senthil Kumar[a]

[a] Centre for Space Science and Technology Education in Asia and the Pacific (CSSTEAP), Dehradun 248001, India
[b] Indian Institute of Remote Sensing, Indian Space Research Organisation, Dept. of Space, Govt. of India, Dehradun 248001, India
[c] Remote Sensing Application Department, Space Technology Institute (STI), Vietnam Academy of Science and Technology (VAST), Viet Nam
[d] International Center for Agricultural Research in the Dry Areas, New Delhi, India

ARTICLE INFO

Keywords:
Forest biomass
Sentinel-2
Spectral variables
Texture variables
Variable optimization
Random Forest

ABSTRACT

Forest biomass is one of the key measurement for carbon budget accounting, carbon flux monitoring, and climate change studies. Hence, it is essential to develop a credible approach to estimate forest biomass and carbon stocks. Our study applied Sentinel-2 satellite imagery combined with field-measured biomass using Random Forest (RF), a machine learning regression algorithm, to estimate forest aboveground biomass (AGB) in Yok Don National Park, Vietnam. A total of 132 spectral and texture variables were extracted from Sentinel-2 imagery (February 7, 2017) to predict AGB of the National Park using RF algorithm. It was found that a combination of 132 spectral and texture variables could predict AGB with an R^2 value of 0.94, RMSE of 34.5 Mgha^{-1} and % RMSE of 18.3%. RF regression algorithm was further used to reduce the number of variables in such a way that a minimum number of selected variables can be able to estimate AGB at a satisfactory level. A combination of 11 spectral and texture variables was identified based on out-of-bag (OOB) estimation to develop an easy-to-use model for estimating AGB. On validation, the model developed with 11 variables was able to predict AGB with R^2 = 0.81, RMSE = 36.67 Mg ha^{-1} and %RMSE of 19.55%. The results found in the present study demonstrated that Sentinel-2 imagery in conjunction with RF-based regression algorithm has the potential to effectively predict the spatial distribution of forest AGB with adequate accuracy.

1. Introduction

Forests sequester a large amount of carbon and play a crucial role in the global climate system (Pan et al., 2013). Quantification of forest biomass is, thus, vital for carbon budget accounting, carbon flux monitoring and for understanding the forest ecosystem response to climate change (Nandy et al., 2019). However, reforestation, afforestation and avoiding deforestation are the mechanisms of tackling climate change (Hunt, 2009; Luong et al., 2015). Therefore, estimation of the forest biomass/carbon stocks not only contributes in Reducing Emissions from Deforestation and forest Degradation (REDD) but also in the sustainable management of forest (Hussin et al., 2014).

Remotely sensed data integrated with forest inventories has become an effective approach to estimate aboveground biomass (AGB) and ultimately carbon stocks. The United Nations collaborative programme on REDD (UN-REDD) has also suggested that national forest monitoring systems should include the use remote sensing (RS) technology for conducting inventory to evaluate forest carbon reference, monitor forest cover, and to assess forest degradation. RS based studies relate image reflectance, spectral indices and image texture with in-situ measurements to estimate biomass (Kushwaha et al., 2014; Lu, 2005; Nelson et al., 2000; Sales et al., 2007; Yadav and Nandy, 2015). With the development of new sensors, improved spatial, spectral, radiometric, and temporal resolutions, RS images can further contribute to fine-scale mapping and frequent monitoring of forest biomass/carbon. Better data integration approaches are also required for accurate and spatially explicit estimations of forest AGB (Nandy et al., 2019).

Lately, machine-learning algorithms including Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF) are increasingly being used to estimate biomass by integrating RS and field data (Dhanda et al., 2017; Nandy et al., 2017; Pandit et al., 2018; Wu et al., 2016). Among a variety of machine learning techniques, the RF algorithm (Breiman, 2001) has been regarded as one of the best methods for classification and regression due to its advantages such

* Corresponding author.
E-mail address: subrato.nandy@gmail.com (S. Nandy).

https://doi.org/10.1016/j.ecoinf.2018.12.010
Received 15 July 2018; Received in revised form 28 December 2018; Accepted 30 December 2018
Available online 31 December 2018

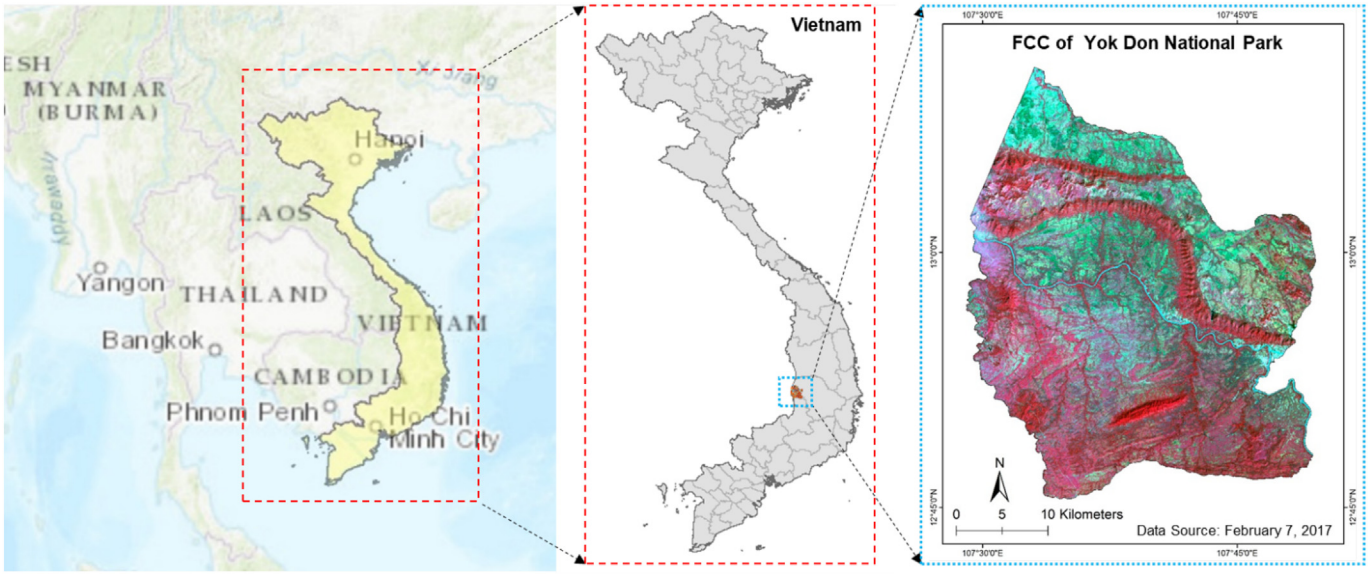

Fig. 1. Location of Yok Don National Park in Vietnam.

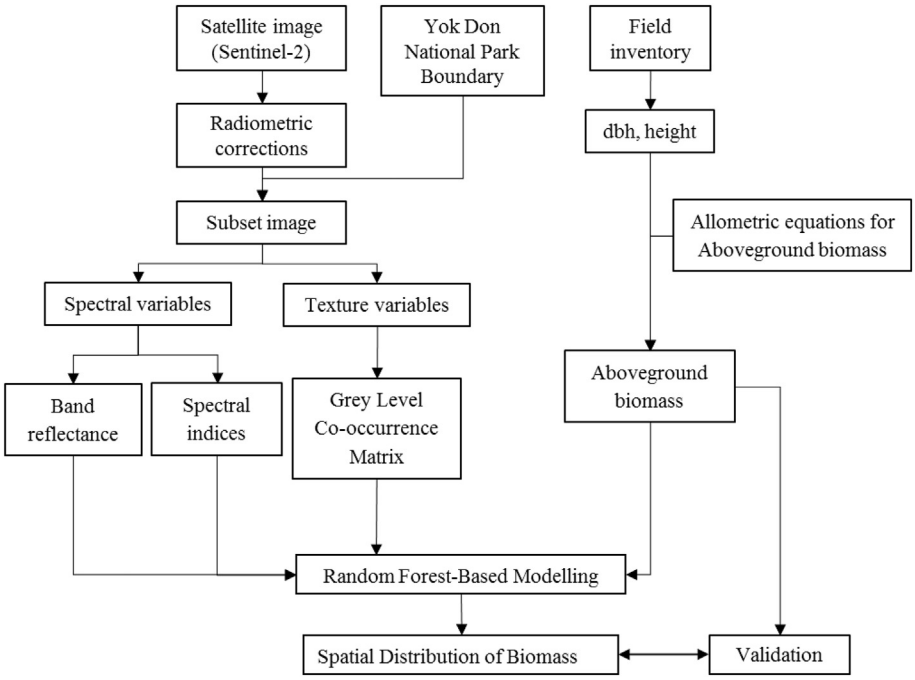

Fig. 2. The methodology for estimation of aboveground biomass in Yok Don National Park, Vietnam.

as high accuracy for estimation outcomes, high speed of computation, robustness and capacity to predict the important variables (Belgiu and Drăguţ, 2016; Cutler et al., 2007). Various studies have used RF algorithm for growing stock and biomass estimation (Chrysafis et al., 2017; Dhanda et al., 2017; Mutanga et al., 2012; Pandit et al., 2018; Pham and Brabyn, 2017).

European Space Agency (ESA) launched polar-orbiting Sentinel-2 sensors with 13 spectral bands in June 2015. Sentinel-2 mission provides a noteworthy improvement in terms of spatial resolution, temporal frequency and spectral diversity over Landsat 8 (Gómez, 2017). The spectral configuration of Sentinel-2 is unique as it comprises red-edge and shortwave infrared (SWIR) bands along with the standards bands available with other commercial and freely available optical satellite data. Furthermore, the data is freely available. Hence, there is a huge potential to use Sentinel-2 data sets for land surface monitoring (Forkuor et al., 2017), assessment of biophysical characteristics (Chrysafis et al., 2019; Pandit et al., 2018), and chemical properties (Ramoelo et al., 2015) at fine scale (at 10 m and 20 m).

Forests cover approximately 40% of the total land area of Vietnam (Vogelmann et al., 2017) which suggests the presence of a large terrestrial carbon stock. However, there are very few studies which demonstrated the potential of RS-based approaches for AGB estimation in Vietnam (Luong et al., 2015; Pham and Brabyn, 2017). Hence, there is an urgent need to develop robust approaches for biomass/carbon stock estimation in the country. The present study aims to estimate forest AGB by integrating RS and field inventory data using RF, a machine-learning algorithm. This approach can contribute to fill knowledge gaps by providing a scientific basis for estimating forest AGB for the present study area and also develop a credible approach which can be used to map and monitor the carbon stock of the country.

Table 1
Spectral and texture variables, extracted from Sentinel-2 satellite imagery, used for developing aboveground biomass prediction models.

S. No.	Independent variables	Details		References
1.	Sentinel 2 Band Reflectance	B2	Blue, 490 nm	Sentinel-2 User Handbook, 2015
		B3	Green, 560 nm	
		B4	Red, 665 nm	
		B5	Red edge1, 705 nm	
		B6	Red edge2, 749 nm	
		B7	Red edge3, 783 nm	
		B8	Near Infra-Red, 842 nm	
		B8a	Near Infra-Red, 865 nm	
		B11	Short Wave Infra-Red, 1610 nm	
		B12	Short Wave Infra-Red, 2190 nm	
2.	Spectral Indices	Atmospherically Resistant Vegetation Index (ARVI)		Kaufman and Tanre, 1992
		Green Chrolophyll Index (CI_g))		Gitelson et al., 2003
		Red Edge Chrolophyll Index (CI_RE)		Gitelson et al., 2003
		Difference Vegetation Index (DVI)		Tucker, 1979
		Enhanced Vegetation Index (EVI)		Huete et al., 2002
		Red Edge 1 Enhanced Vegetation Index (EVI_RE1)		Chrysafis et al., 2017
		Red Edge 2 Enhanced Vegetation Index (EVI_RE2)		Chrysafis et al., 2017
		Red Edge 3 Enhanced Vegetation Index (EVI_RE3)		Chrysafis et al., 2017
		NIR2 Enhanced Vegetation Index (EVI_NIR2)		Chrysafis et al., 2017
		Green Atmospherically Resistant Index (GARI)		Gitelson et al., 1996
		Green Difference Vegetation Index (GDVI)		Sripada et al., 2006
		Green Normalized Difference Vegetation Index (GNDVI)		Gitelson and Merzlyak, 1998
		Inverted Red Edge Chlorophyll Index (IRECI)		Frampton et al., 2013
		Modified Chlorophyll Absorption Index(MCARI)		Haboudane et al., 2004
		Modified Soil Adjusted Vegetation Index (MSAVI)		Qi et al., 1994
		Moisture Stress Index (MSI)		Hunt and Rock, 1989
		Modified Simple Ratio (MSR)		Chen, 1996
		Red Edge1 Modified Simple Ratio(MSR_RE1)		Chrysafis et al., 2017
		Red Edge2 Modified Simple Ratio(MSR_RE2)		Chrysafis et al., 2017
		Red Edge3 Modified Simple Ratio(MSR_RE3)		Chrysafis et al., 2017
		NIR2 Modified Simple Ratio (MSR_NIR2)		Chrysafis et al., 2017
		Normalized Difference Infrared Index (NDII)		Buschmann, 1993
		Normalized Difference Vegetation Index (NDVI)		Rouse Jr et al., 1974
		Red Edge Normalized Difference Vegetation Index (NDVI705)		Gitelson and Merzlyak, 1998
		Normalized Difference Water Index (NDWI)		Gao, 1996
		Normalized Green (NG)		Kender, 1976
		Non-Linear Index (NLI)		Goel and Qin, 1994
		Red Edge1 Non Linear Index (NLI_RE1)		Chrysafis et al., 2017
		Red Edge2 Non Linear Index (NLI_RE2)		Chrysafis et al., 2017
		Red Edge3 Non Linear Index (NLI_RE3)		Chrysafis et al., 2017
		NIR2 Non Linear Index (NLI_NIR2)		Chrysafis et al., 2017
		Normalized Near Infrared (NNIR)		Majasalmi and Rautiainen, 2016
		Normalized Red (NR)		Kender, 1976
		Plant Senescence Reflectance Index (PSRI)		Merzlyak et al., 1999
		NIR Plant Senescence Reflectance Index(PSRI NIR)		Merzlyak et al., 1999
		Pigment Specific Simple Ratio (PSSR)		Blackburn, 1998
		Renormalized Difference Vegetation Index (RDVI)		Roujean and Breon, 1995
		Soil Adjusted Vegetation Index (SAVI)		Huete, 1988
		Transformed Soil Adjusted Vegetation Index (TSAVI)		Baret et al., 1989
		Vegetation Index Green (VARI_g)		Chrysafis et al., 2017
		Wide DynamicRange Vegetation Index (WDRVI)		Gitelson, 2004
		Red Edge Wide Dynamic Range (WDRVI_RE)		Gitelson, 2004; Majasalmi and Rautiainen, 2016
3.	Texture Variables	Gray-level Co-occurrence Matrix with a five by five-pixel window	Mean	Haralick et al., 1973
			Variance	
			Homogeneity	
			Contrast	
			Dissimilarity	
			Entropy	
			Second Moment	
			Correlation	

2. Materials and methods

2.1. Study area

The present study was conducted in Yok Don National Park, which is located in the Central Highlands of Vietnam, between 12°45′00″ to 13°10′00″N, and 107°29′00″ to 107°48′00″E (Fig. 1). There are two major forest types in the study area: (i) evergreen broadleaved forest with dominant species, viz., *Michelia mediocris, Cinamomum iners, Syzygium zeylanicum, S. wightianum, Garruga pierrei, Gonocaryum lobbianum, Schima superba, Camellia assamica,* and *Lithocarpus fenestratus*; and (ii) deciduous broadleaved forest with the dominant tree species including *Dipterocarpus tuberculatus, D. obtusifolius, Terminalia tomentosa,* and *Shorea obtuse* (Luong et al., 2017). The area experiences a tropical monsoon climate with a dry season (November to April), and typical rainy season (May to October) with the mean annual rainfall of 1540 mm, and the mean monthly temperature of approximately 25 °C (Canh et al., 2009; Luong et al., 2017).

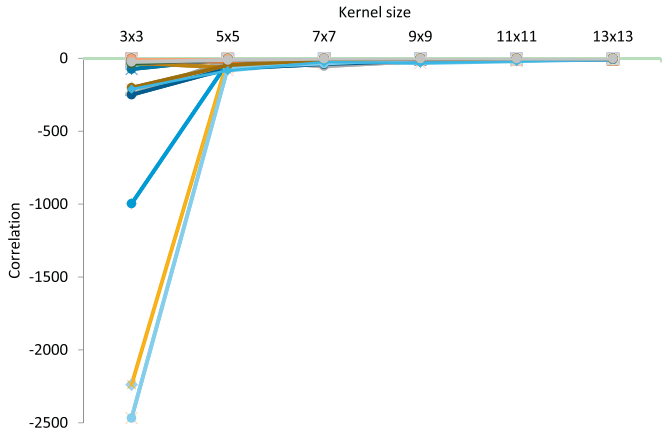

Fig. 3. Values of correlation images at different kernel sizes e.g., 3 × 3, 5 × 5, 7 × 7, 9 × 9, 11 × 11 and 13 × 13 at the field sampling points.

Table 2
Model parameters (M_{try}, N_{tree}) and model accuracy (R^2, RMSE and %RMSE) of three Random Forest based models used in the present study.

Model	No. of variables	M_{try}	N_{tree}	R^2	RMSE (Mg ha^{-1})	%RMSE
Model 1 (Spectral)	52	25	500	0.92	36.0	19.3
Model 2 (Texture)	80	39	500	0.93	35.9	19.1
Model 3 (Spectral + Texture)	132	66	500	0.94	34.5	18.3

2.2. Methodology

The purpose of the present study is to estimate the forest AGB by integrating RS and field inventory data using RF regression algorithm. The overall methodology is shown in Fig. 2.

2.2.1. Satellite data processing and variable extraction

The present study used Sentinel-2 satellite imagery of 7 February 2017. Sentinel-2 provides optical satellite data in 13 spectral bands, from the visible and the near-infrared to the shortwave infrared. In the present study 10 spectral bands, excluding bands 1 (Coastal aerosol), 9 (Water vapour) and 10 (SWIR-Cirrus) were used (Table 1).

Orthorectified satellite data was downloaded (https://earthexplorer.usgs.gov/) and a subset image was extracted from the acquired reflectance image using the study area boundary. The reflectance image of the study area was used to extract: (i) spectral variables including 10 bands reflectance, and 42 spectral indices; and (ii) texture variables (Haralick et al., 1973) (mean, contrast, correlation, entropy, dissimilarity, homogeneity, variance, and second angular moment) (Table 1). Gray Level Co-occurrence Matrix (GLCM) technique for texture analysis (Haralick et al., 1973) was applied to extract various texture variables. Correlation texture variable was used to identify the optimum kernel size for texture variables extraction. The results showed that there was significant variation in correlation variable for 3 × 3 kernel size while the correlation almost converged at 5 × 5 kernel size (Fig. 3) and hence 5 × 5 was chosen as optimum kernel size for extracting all the texture variables.

2.2.2. Sample design and field data collection

For this study, 58 sample plots of 500 m^2 (25 m × 20 m) for deciduous broadleaved forest and 1000 m^2 (33.3 m × 30 m) for evergreen broadleaved forest were laid. Sample plots were laid in such a way that forest diversity and topographical variation can be captured as suggested by Luong et al. (2015). Out of the 58 plots, 39 plots were selected randomly for RF modelling and remaining 19 plots were employed for validation of the model. At each sample plot, GPS coordinates (accuracy < 5 m) were recorded along with tree species name, diameter at breast height (dbh) and tree height. The field data was collected during April 2014.

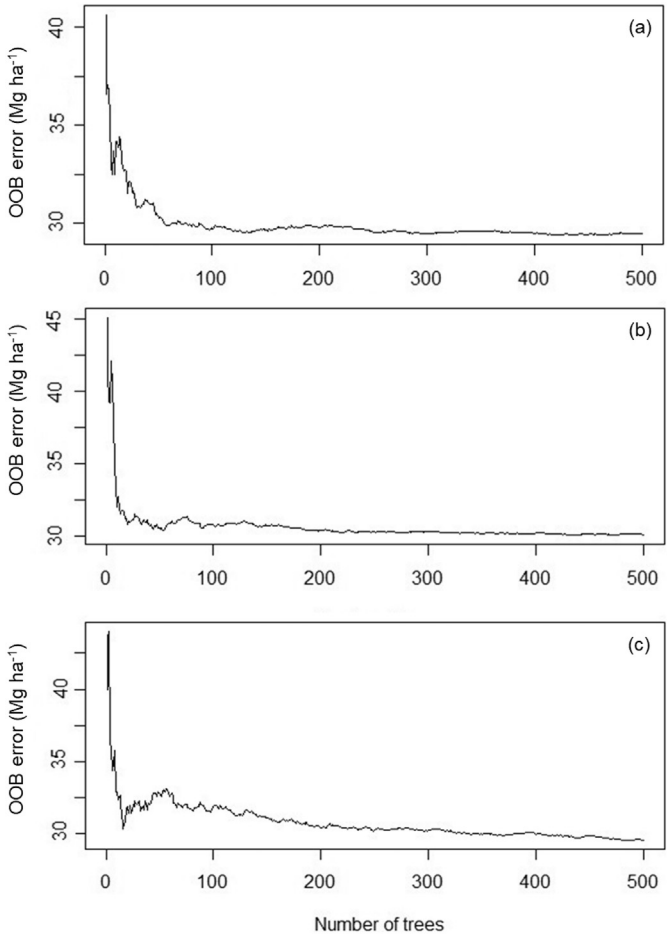

Fig. 4. Out-of-bag (OOB) error versus number of trees in Random Forest for three models: (a) Model 1: Considering only spectral variables; (b) Model 2: Considering only texture variables; (c) Model 3: Considering both spectral and texture variables.

The present study utilises allometric equations for AGB estimation of evergreen broadleaved and deciduous broadleaved forest, which were developed by the UN-REDD program in 2012 for highlands of Vietnam (Luong et al., 2015). The optimized equations for biomass prediction at tree level are:

For evergreen broadleaved forests:

$$AGB = 0.0530 \times (D^2 \times H^{0.7})^{1.0072} \quad (1)$$

For deciduous broadleaved forests:

$$AGB = 0.0154 \times (D^2 \times H^{0.7})^{1.1682} \quad (2)$$

where, D = dbh; H = height of tree.

2.2.3. Estimating AGB using random forest

The *randomForest* package in R was used in the present study. RF regression algorithm was used to determine the optimal independent variables, including spectral and texture variables, with respect to the dependent variable- the field-measured biomass. Both spectral (vegetation indices and band reflectances) and texture values were extracted at the field-measured biomass plot locations. Two input parameters are needed to be optimized in RF, viz., N_{tree}, the number of regression trees grown based on a bootstrap sample of the observations and M_{try}, the number of variables fed to each predictor tree, its default value used by the algorithm is $\log_2(M + 1)$, where M is the number of observations.

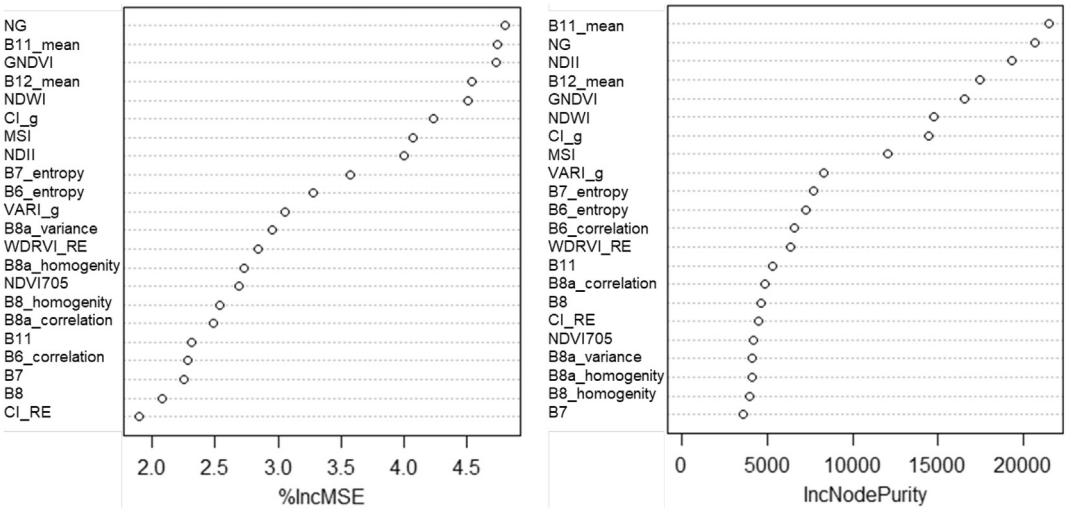

Fig. 5. Random Forest predicted variable importance for Model 3 showing top variables for aboveground biomass prediction.

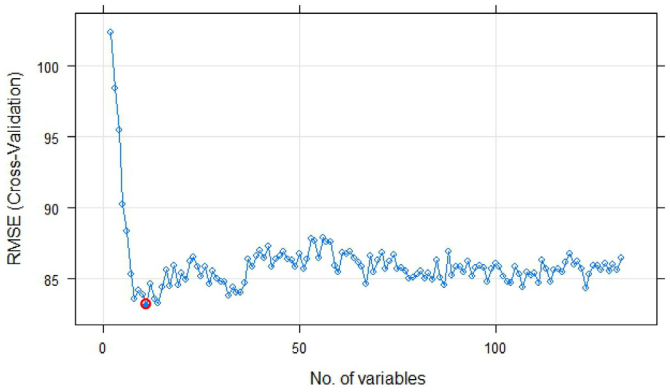

Fig. 6. Selection of optimum number of variables based on the least RMSE for 10-fold cross-validation.

About two-thirds of the samples (in-bag samples) are used to train the trees and remaining one third (out-of-bag (OOB) samples) are used in an internal cross-validation technique for estimating the OOB error (Belgiu and Drăguţ, 2016). Selecting optimum values of M_{try} and N_{tree} is essential to construct a RF model with low OOB error. The algorithm was repetitively run with various values of M_{try} and N_{tree} to achieve the optimal accuracy. M_{try} values were obtained using the *tuneRF* function of the *randomForest* package. Performing the *tuneRF* function repeatedly, to assign the value of M_{try} with the most frequent occurrence, is an acceptable approach for tackling variation of M_{try} (Pham and Brabyn, 2017).

To estimate biomass, three RF-based models were investigated using different combinations of variables: (i) Model 1 – Predicted AGB using only spectral variables (52 nos.), (ii) Model 2 - Predicted AGB using only texture variables (80 nos.) and (iii) Model 3 - Predicted AGB using both spectral and texture variables (132 nos.). The R^2, RMSE, and % RMSE of these models were compared to select the final model. To optimize the number of variables, a recursive feature elimination integrated with RF regression was used (Mutanga et al., 2012). The method involves the selection of variables by progressively eliminating the least promising variables and which is further evaluated using cross-validation as the number of predictors is reduced.

2.2.4. Uncertainty mapping

Uncertainty in predictions of the RF algorithm (regression-type) can be assessed through different techniques such as Jackknife-after-Bootstrap, U-statistics, Monte Carlo simulations, Quantile Regression Forests. Since RF is a non-parametric ensemble method, there is no direct quantification of prediction error like in traditional regression approaches (Coulston et al., 2016). Here, Monte Carlo simulation approach was used to quantify prediction uncertainty. Monte Carlo simulation uses samples of the variables according to their probabilistic characteristics and then feeds them into the performance function.

3. Results

3.1. AGB estimation using RF

RF algorithm was run repeatedly to obtain the optimum M_{try} and N_{tree} values (Table 2). N_{tree} of 500 was chosen for all the three models as OOB error of prediction was found to be the lowest (Fig. 4). For Model 1, the relationship between observed and predicted biomass illustrated an R^2 value of 0.92 and %RMSE of 19.3% whereas Model 2 depicted an R^2 of 0.93 and %RMSE of 19.1% and for Model 3, R^2 was 0.94 and % RMSE was 18.3% (Table 2). Hence, Model 3 was used to predict the AGB of the study area as it provided the highest accuracy.

When the RF was repetitively performed with various inputs parameters to obtain variable importance for Model 3, the mutual rankings of the variables altered, but a set of top variables remained unchanged (Fig. 5). To optimize the number of variables for Model 3, cross-validation process with recursive elimination was used. The optimum number of variables based on least RMSE for 10-fold cross-validation was found to be eleven (Fig. 6). The top 11 variables, viz., B11 mean, Normalized Green (NG), Normalized Difference Infrared Index (NDII), B12 mean, Green Normalized Difference Vegetation Index (GNDVI), Normalized Difference Water Index (NDWI), Green Chrolophyll Index (CI_g), Moisture Stress Index (MSI), Vegetation Index Green (VARI_g), B7 entropy, and B6 entropy, were used in the final model to predict the AGB.

3.2. Spatial pattern of AGB

The spatial distribution of AGB (Fig. 7a) was generated at 20 m spatial resolution using the RF algorithm with 11 spectral and texture variables. The values of AGB ranged from 77.33 to 381.78 Mg ha^{-1}. The AGB map obtained using the RF algorithm was validated using 19 validation plots. On validation, it was found that RF algorithm was able to predict AGB with R^2 of 0.81 (Fig. 8), RMSE of 36.67 Mg ha^{-1} and % RMSE of 19.55%.

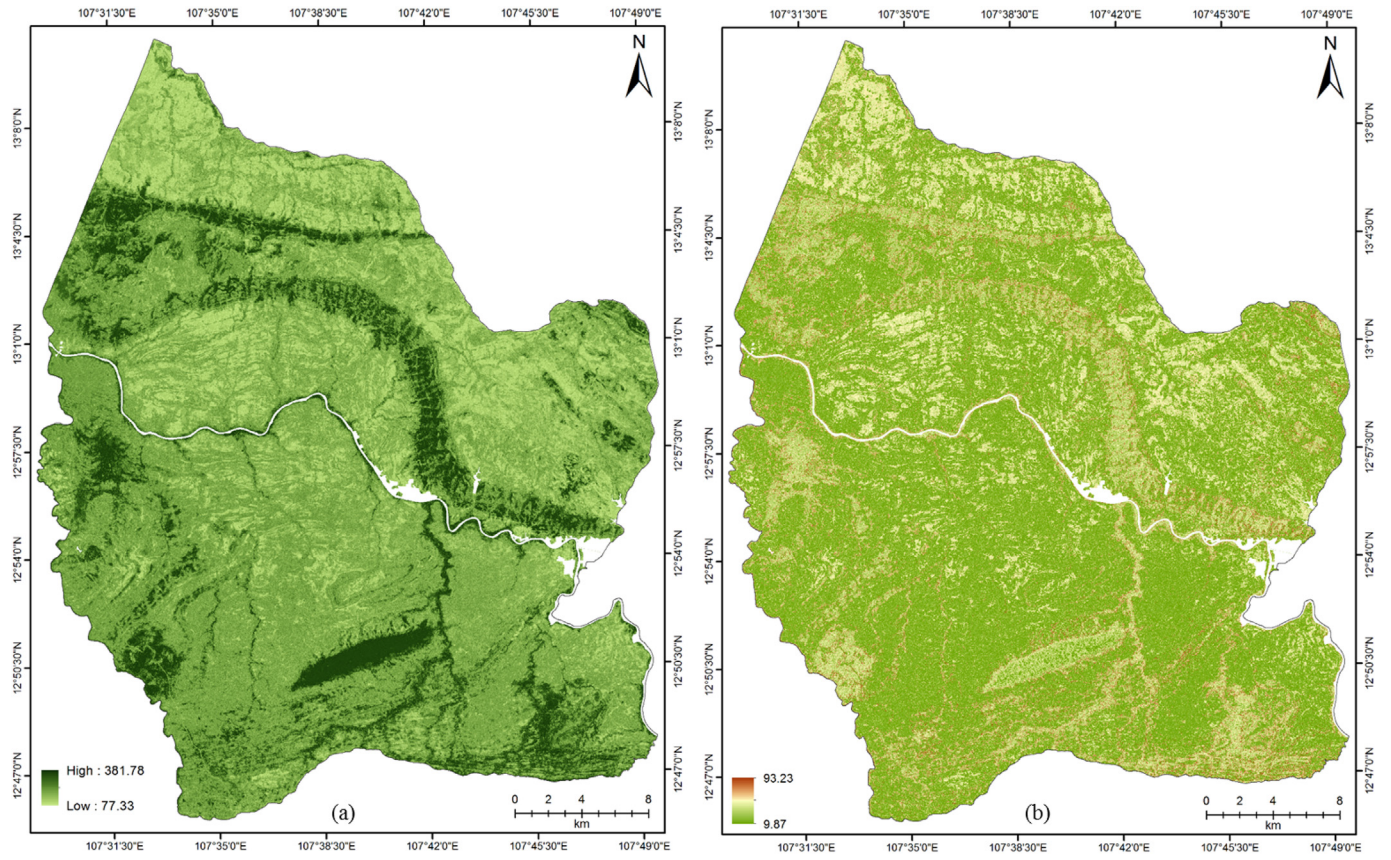

Fig. 7. (a) Spatial distribution of aboveground biomass of Yok Don National Park, Vietnam, (b) Uncertainty map of aboveground biomass of Yok Don National Park, Vietnam.

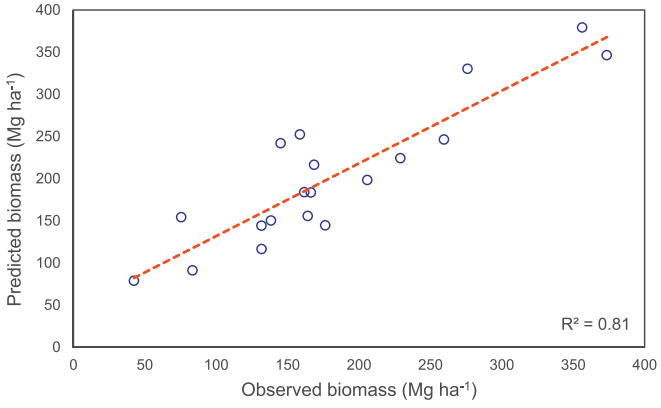

Fig. 8. Observed vs. predicted aboveground biomass for validation plots.

3.3. Uncertainty map of AGB

The uncertainty of biomass estimation was shown in Fig. 7b. The average value of uncertainty was 31.99 Mg ha^{-1} within the range between 9.87 and 93.27 Mg ha^{-1}. Percentage of mean uncertainty with respect to mean predicted biomass was 16.73%.

4. Discussion

In the present study, the spatial distribution of AGB was predicted with R^2 of 0.81, RMSE of 36.67 Mg ha^{-1} and %RMSE of 19.55% using RF regression algorithm. Table 3 summarises the findings of few recent studies where machine learning algorithms were used to estimate forest AGB in different ecological settings. It has been observed that machine learning algorithms, like ANN, RF, SVM, etc., were used in different climatic conditions to estimate the forest AGB with considerable accuracy. This demonstrates the capability of machine learning techniques, like RF, in estimating AGB with substantial accuracy compared to linear modelling (Dube and Mutanga, 2015). The non-parametric algorithms have the ability to efficiently address the non-linear relationship between forest AGB and RS data (Liu et al., 2017). The capability to handle non-linearity and quantify the importance of independent variables makes RF an effective algorithm (Pandit et al., 2018).

In the present study, it was observed that a mixture of spectral and texture variables yielded higher accuracy for biomass estimation. Pham and Brabyn (2017) also found that a combination of spectral and texture variables yielded a strong relationship (R^2 = 0.73 and RMSE = 78.2 Mg ha^{-1}) with the AGB of mangrove forest of Cangio district, Vietnam. Pandit et al. (2018) examined the potential of Sentinel-2 for estimating forest AGB using RF algorithm. They found that the 20 important top spectral variables derived from the Sentinel-2 satellite data could explain AGB with high accuracy (R^2 = 0.81 and RMSE = 25.32 Mg ha^{-1}). Whereas, in the present study, 11 important top variables showed strong relationship with the forest AGB. The top variables mostly comprised of SWIR and red-edge band based variables. This highlights the utility of Sentinel-2 data having better spectral, with additional SWIR and red-edge bands, and spatial resolutions among the available medium resolution sensors. This observation also corroborates with the findings of Pandit et al. (2018). In another study, in-situ data obtained by Relasphone was integrated with optical remotely sensed data to generate growing stock volume and AGB maps in Mexico (Molinier et al., 2016). The results demonstrated that green and SWIR bands were suitable for estimating AGB. SWIR band has shown stronger relation with field measured biomass irrespective of data and

Table 3
Studies on forest biomass estimation using machine learning algorithms in different ecological settings.

S. No.	Modelling approach	Satellite data	Location	Ecological settings	Model performance		Reference
					R^2	RMSE (Mg ha^{-1})	
1.	Artificial Neural Network (ANN)	Resourcesat-1 LISS-III	Barkot forest, Uttarakhand, India	Tropical Moist Deciduous forest, subtropical climate	0.74	93.41	Nandy et al., 2017
2.	Support Vector Regression (SVR)	Landsat 7 ETM+	Zhejiang Province, China	Subtropical monsoon climate	0.38	34.61	Wu et al., 2016
		ICESat/GLAS and WorldView-2	Doon valley, Uttarakhand, India	Tropical Moist Deciduous forest, subtropical climate	0.89	13.60	Dhanda et al., 2017
3.	Random Forest (RF)	Landsat 7 ETM+	Zhejiang Province, China	Subtropical monsoon climate	0.63	26.22	Wu et al., 2016
		RapidEye	KwaZulu-Natal Province, South Africa	Plantation forest, subtropical climate	0.37	59.27	Dube et al., 2014
		ICESat/GLAS and WorldView-2	Doon valley, Uttarakhand, India	Tropical Moist Deciduous forest, subtropical climate	0.84	20.57	Dhanda et al., 2017
		WorldView-2	ISimangaliso Wetland Park, KwaZulu-Natal Province, South Africa	Wetland vegetation, subtropical climate	–	4.41	Mutanga et al., 2012
		SPOT 4, SPOT 5	Cangio Mangrove Forest, Cangio, Ho Chi Minh City, Vietnam	Mangrove forest, subtropical climate	0.73	78.20	Pham and Brabyn, 2017
		Sentinel-2	Evros prefecture, Rhodopes mountain range, Greece	Mediterranean forest	0.63	63.11[a]	Chrysafis et al., 2017
		Sentinel-2	Parsa National Park, Nepal	Central-Southern part of Nepal, Subtropical climate	0.81	25.32	Pandit et al., 2018
		Sentinel-2	Yok Don National Park, Vietnam	Central Highlands of Vietnam, tropical monsoon climate	0.81	36.67	Present study
4.	Stochastic Gradient Boosting (SGB)	Landsat 7 ETM+	Zhejiang Province, China	Subtropical monsoon climate	0.55	28.64	Wu et al., 2016
		RapidEye	KwaZulu-Natal Province, South Africa	Plantation forest, subtropical climate	0.61	43.39	Dube et al., 2014

[a] $m^3 ha^{-1}$ (growing stock volume).

environmental settings (Chrysafis et al., 2017; Nandy et al., 2017; Yadav and Nandy, 2015). Vegetation reflects maximum energy at NIR region of the electromagnetic spectrum, but unable to provide any information of soil under the vegetation, while SWIR band can differentiate the moisture content of vegetation and soil (Brown et al., 2016). Hence, SWIR band could capture the vegetation cover with underneath soil condition more efficiently.

Our study found that vegetation indices, viz. NDII and NG also had strong relationship with the AGB. The results are comparable with the findings of the previous studies. Pham and Brabyn (2017) in a study conducted in the mangrove forest in Cangio district of Vietnam found that NDII as one of the top variables which has a strong relationship with AGB. In their study, they considered 123 spectral and texture variables to estimate AGB. Majasalmi and Rautiainen (2016), while estimating the canopy biophysical properties in boreal forests of Finland found that among the vegetation indices, NG had the strongest relationship with AGB. Adam et al. (2014) stated that vegetation indices have the capability to reduce the impacts of shadows and environmental conditions on reflectance and hence can improve their relationship with the AGB.

There are studies that have used RF for prediction of different biophysical parameters (Baccini et al., 2008; Dhanda et al., 2017; Powell et al., 2010). However, uncertainties of those predictions are mapped in very few cases (Coulston et al., 2016). Quantifying prediction uncertainty is important to evaluate model performance, specifically when the resultant prediction is used as an input parameter to other model. In the present study, uncertainty map of biomass gives a clear picture about the spatial distribution of uncertainty. Uncertainty in dense forest area was less and higher values were found in sparse vegetative area. This implies that, heterogeneity of vegetative surface may increase uncertainty of the model.

5. Conclusions

The present study demonstrates an approach for forest biomass estimation by integrating remotely sensed data with field inventory data with the aid of a machine learning algorithm. RF was applied to identify the dominant spectral and texture variables to estimate forest AGB. Using a combination of 11 spectral and texture variables, the RF algorithm effectively predicted the spatial distribution of forest AGB of Yok Don National Park, Vietnam.

The method provides a scientific basis for estimating high-resolution forest AGB integrating field inventory and Sentinel-2 sensor data with the help of machine learning technique. The methodology can be adopted for mapping and monitoring the forest biomass/carbon stock of Vietnam. Being time and cost-effective, this approach can be utilized for studying the spatiotemporal changes in biomass. The estimated spatial distribution of AGB is not only a crucial source for monitoring biomass/carbon stocks, but also greatly contributes to determining a national reference scenario of biomass and helps to support national REDD strategy implementation in Vietnam.

Acknowledgements

The first author gratefully acknowledges Centre for Space Science and Technology Education in Asia and the Pacific (CSSTEAP) for financial support during the study. The authors are thankful to the Director, Indian Institute of Remote Sensing, ISRO, Dehradun for his support during the study. The authors are thankful to VT-UD.05/17-20 project of Vietnam for providing the field data. Thanks are also due to the anonymous reviewers for the critical review of the manuscript.

References

Adam, E., Mutanga, O., Abdel-Rahman, E.M., Ismail, R., 2014. Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: exploratory of in situ hyperspectral indices and random forest regression. Int. J. Remote Sens. 35, 693–714.
Baccini, A., Laporte, N., Goetz, S.J., Sun, M., Dong, H., 2008. A first map of tropical Africa's above-ground biomass derived from satellite imagery. Environ. Res. Lett. 3 (4), 045011.
Baret, F., Guyot, G., Major, D.J., 1989. TSAVI: a vegetation index which minimizes soil brightness effects on LAI and APAR estimation. In: IEEE International Geoscience and Remote Sensing Symposium and 12th Canadian Symposium on Remote Sensing, Vancouver. vol. 3. pp. 1355–1358.
Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31.
Blackburn, G.A., 1998. Quantifying chlorophylls and caroteniods at leaf and canopy scales: An evaluation of some hyperspectral approaches. Remote Sens. Environ. 66 (3), 273–285.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Brown, D., Jorgenson, M.T., Kielland, K., Verbyla, D.L., Prakash, A., Koch, J.C., 2016. Landscape effects of wildfire on permafrost distribution in interior Alaska derived from remote sensing. Remote Sens. 8 (8), 654.
Buschmann, C., 1993. Fernerkundung von Pflanzen. Naturwissenschaften 80 (10), 439–453.
Canh, N.X., Quynh, H.Q., Anh, L.T., Luong, N.V., 2009. Report on Conservation Planning and Sustainable Development of Yok Don National Park in 2010–2020. Ministry of Agriculture and Rural Development of Vietnam, Hanoi, Vietnam.
Chen, J.M., 1996. Evaluation of vegetation indices and a modified simple ratio for boreal applications. Can. J. Remote Sens. 22 (3), 229–242.
Chrysafis, I., Mallinis, G., Siachalou, S., Patias, P., 2017. Assessing the relationships between growing stock volume and Sentinel-2 imagery in a Mediterranean forest ecosystem. Remote Sens. Lett. 8 (6), 508–517.
Chrysafis, I., Mallinis, G., Tsakiri, M., Patias, P., 2019. Evaluation of single-date and multi-seasonal spatial and spectral information of Sentinel-2 imagery to assess growing stock volume of a Mediterranean forest. Int. J. Appl. Earth Observ. Geoinform. 77, 1–14.
Coulston, J.W., Blinn, C.E., Thomas, V.A., Wynne, R.H., 2016. Approximating prediction uncertainty for random forest regression models. Photogramm. Eng. Remote. Sens. 82 (3), 189–197.
Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecol. 88 (11), 2783–2792.
Dhanda, P., Nandy, S., Kushwaha, S.P.S., Ghosh, S., Krishna Murthy, Y.V.N., Dadhwal, V.K., 2017. Optimising spaceborne LiDAR and very high resolution optical sensor parameters for biomass estimation at ICESat/GLAS footprint level using regression algorithms. Prog. Phys. Geogr. 41 (3), 247–267.
Dube, T., Mutanga, O., 2015. Evaluating the utility of the medium-spatial resolution Landsat 8 multi-spectral sensor in quantifying aboveground biomass in Umgeni catchment, South Africa. ISPRS J. Photogramm. Remote Sens. 101, 36–46.
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 advanced machine learning algorithms. Sensors 14 (8), 15348–15370.
Forkuor, G.F., Dimobe, K., Serme, I., Tondoh, J.E., 2017. Landsat-8 vs. Sentinel-2: examining the added value of Sentinel-2's red-edge bands to land-use and land cover mapping in Burkina Faso. GISci. Remote Sens. 1–24.
Frampton, W.J., Dash, J., Watmough, G., Milton, E.J., 2013. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. ISPRS J. Photogramm. Remote Sens. 82, 83–92.
Gao, B.C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58 (3), 257–266.
Gitelson, A.A., 2004. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. J. Plant Physiology 161 (2), 165–173.
Gitelson, A.A., Merzlyak, M.N., 1998. Remote sensing of chlorophyll concentration in higher plant leaves. Adv. Space Res. 22 (5), 689–692.
Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ. 58 (3), 289–298.
Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. J. Plant Physiol. 160 (3), 271–282.
Goel, N.S., Qin, W., 1994. Influences of canopy architecture on relationships between various vegetation indices and LAI and FPAR: a computer simulation. Remote Sens. Rev. 10 (4), 309–347.
Gómez, M.G.C., 2017. Joint Use of Sentinel-1 and Sentinel-2 for Land Cover Classification: A Machine Learning Approach. Master's Thesis. Lund University, Lund, Sweden.
Haboudane, D., Miller, J.R., Pattey, E., Zarco-Tejada, P.J., Strachan, I.B., 2004. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: modeling and validation in the context of precision agriculture. Remote Sens. Environ. 90 (3), 337–352.
Haralick, R.M., Shanmugam, K., Denstien, I., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. 3, 610–621.
Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25 (3), 295–309.
Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83 (1), 195–213.
Hunt, C.A., 2009. Carbon Sinks and Climate Change: Forests in the Fight against Global Warming. Edward Elgar Publishing.
Hunt, E.R., Rock, B.N., 1989. Detection of changes in leaf water content using near-and middle-infrared reflectances. Remote Sens. Environ. 30 (1), 43–54.

Hussin, Y.A., Gilani, H., van Leeuwen, L., Murthy, M., Shah, R., Baral, S., Tsendbazar, N.E., Shrestha, S., Shah, S.K., Qamer, F.M., 2014. Evaluation of object-based image analysis techniques on very high-resolution satellite image for biomass estimation in a watershed of hilly forest of Nepal. Appl. Geomatics 6 (1), 59–68.
Kaufman, Y.J., Tanre, D., 1992. Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. IEEE Trans. Geosci. Remote Sens. 30 (2), 261–270.
Kender, J.R., 1976. Saturation, Heu, and Normalized Color: Calculation, Digitization Effects, and Use. Carnegie-Mellon Univ Pittsburgh Pa Dept. of Computer Science.
Kushwaha, S.P.S., Nandy, S., Gupta, M., 2014. Growing stock and woody biomass assessment in Asola-Bhatti Wildlife Sanctuary, Delhi, India. Environ. Monit. Assess. 186 (9), 5911–5920.
Liu, K., Wang, J., Zeng, W., Song, J., 2017. Comparison and evaluation of three models for estimating forest aboveground biomass using TM and GLAS data. Remote Sens. 9, 341.
Lu, D., 2005. Aboveground biomass estimation using Landsat TM data in the Brazilian Amazon. Int. J. Remote Sens. 26, 2509–2525.
Luong, N.V., Tateishi, R., Hoan, N.T., Tu, T.T., 2015. Forest change and its effect on biomass in Yok Don National Park in central highlands of Vietnam using ground data and geospatial techniques. Adv. Remote Sens. 4 (02), 108.
Luong, N.V., Tateishi, R., Kondoh, A., Anh, N.D., Hoan, N.T., 2017. Land cover mapping in Yok Don National Park, Central Highlands of Viet Nam using Landsat 8 OLI images. Vietnam J. Earth Sci. 39 (4), 393–406.
Majasalmi, T., Rautiainen, M., 2016. The potential of Sentinel-2 data for estimating biophysical variables in a boreal forest: a simulation study. Remote Sens. Lett. 7 (5), 427–436.
Merzlyak, M.N., Gitelson, A.A., Chivkunova, O.B., Rakitin, V.Y., 1999. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. Physiol. Plant. 106 (1), 135–141.
Molinier, M., López-Sánchez, C.A., Toivanen, T., Korpela, I., Corral-Rivas, J.J., Tergujeff, R., Häme, T., 2016. Relasphone—Mobile and participative in situ forest biomass measurements supporting satellite image mapping. Remote Sens. 8 (10), 869.
Mutanga, O., Adam, E., Cho, M.A., 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int. J. Appl. Earth Obs. Geoinformation 18, 399–406.
Nandy, S., Singh, R., Ghosh, S., Watham, T., Kushwaha, S.P.S., Kumar, A.S., Dadhwal, V.K., 2017. Neural network-based modelling for forest biomass assessment. Carbon Manag. 8 (4), 305–317.
Nandy, S., Ghosh, S., Kushwaha, S.P.S., Kumar, A.S., 2019. Remote sensing-based forest biomass assessment in northwest Himalayan landscape. In: Navalgund, R.R., Senthil Kumar, A., Nandy, S. (Eds.), Remote Sensing of Northwest Himalayan Ecosystems. Springer, Singapore, pp. 285–311.
Nelson, R.F., Kimes, D.S., Salas, W.A., Routhier, M., 2000. Secondary forest age and tropical forest biomass estimation using Thematic Mapper imagery. Biogeosciences 50, 419–431.
Pan, Y., Birdsey, R.A., Phillips, O.L., Jackson, R.B., 2013. The structure, distribution, and biomass of the world's forests. Annu. Rev. Ecol. Evol. Syst. 44, 593–622.
Pandit, S., Tsuyuki, S., Dube, T., 2018. Estimating above-ground biomass in sub-tropical buffer zone community Forests, Nepal, using Sentinel 2 data. Remote Sens. 10 (4), 601.
Pham, L.T., Brabyn, L., 2017. Monitoring mangrove biomass change in Vietnam using SPOT images and an object-based approach combined with machine learning algorithms. ISPRS J. Photogramm. Remote Sens. 128, 86–97.
Powell, S.L., Healey, S.P., Cohen, W.B., Kennedy, R.E., Moisen, G.G., Pierce, K.B., Ohmann, J.L., 2010. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: a comparison of empirical modeling approaches. Remote Sens. Environ. 114 (5), 1053–1068.
Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. Remote Sens. Environ. 48 (2), 119–126.
Ramoelo, A., Cho, M., Mathieu, R., Skidmore, A.K., 2015. Potential of Sentinel-2 spectral configuration to assess rangeland quality. J. Appl. Remote Sens. Environ. 124, 516–533.
Roujean, J.L., Breon, F.M., 1995. Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. Remote Sens. Environ. 51 (3), 375–384.
Rouse Jr., J., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring Vegetation Systems in the Great Plains with ERTS.
Sales, M.H., Souza Jr., C.M., Kyriakidis, P.C., Roberts, D.A., Vidal, E., 2007. Improving spatial distribution estimation of forest biomass with geostatistics: a case study for Rondônia, Brazil. Ecol. Model. 205, 221–230.
Sentinel-2 User Handbook, 2015. ESA Standard Document. (Issue 1 Rev 2).
Sripada, R.P., Heiniger, R.W., White, J.G., Meijer, A.D., 2006. Aerial color infrared photography for determining early in-season nitrogen requirements in corn this project was supported in part by initiative for future agriculture and food systems grant no. 00-52103-9644 from the USDA cooperative state research, education, and extension service. Agron. J. 98 (4), 968–977.
Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8 (2), 127–150.
Vogelmann, J.E., Khoa, P.V., Lan, D.X., Shermeyer, J., Shi, H., Wimberly, M.C., Duong, H.T., Huong, L.V., 2017. Assessment of forest degradation in Vietnam using landsat time series data. Forests 8 (7), 238.
Wu, C., Shen, H., Shen, A., Deng, J., Gan, M., Zhu, J., Xu, H., Wang, K., 2016. Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery. J. Appl. Remote. Sens. 10 (3), 035010.
Yadav, B.K.V., Nandy, S., 2015. Mapping aboveground woody biomass using forest inventory, remote sensing and geostatistical techniques. Environ. Monit. Assess. 187 (5), 308.