Evaluation of machine learning methods and multi-source remote sensing data combinations to construct forest above-ground biomass models

Yan, Xingguang; Li, Jing; Smith, Andy; Yang, Di; Ma, Tianyue; Su, Yiting; Shao, Jiahao

International Journal of Digital Earth

DOI:
10.1080/17538947.2023.2270459

Published: 01/11/2023

Evaluation of machine learning methods and multi-source remote sensing data combinations to construct forest above-ground biomass models

Xingguang Yan[a,b,c], Jing Li[a*], Andrew R. Smith[b,c], Di Yang[d], Tianyue Ma[a], YiTing Su[a] and Jiahao Shao[a]

[a]*College of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China;*

[b]*School of Natural Sciences, Bangor University, Bangor, Gwynedd, LL57 2UW, UK*

[c]*Environment Centre Wales, Bangor University, Bangor, Gwynedd, LL57 2UW, UK*

[d]*Wyoming Geographic Information Science Center, University of Wyoming, WY 82070, USA*

*Author for Correspondence: Jing Li, Email: lijing@cumtb.edu.cn

ORCID

Xingguang Yan https://orcid.org/0009-0001-8280-4568

Andrew R. Smith http://orchid.org/0000-0001-8580-278X

Di Yang http://orchid.org/0000-0002-4010-6163

Abstract: Rapid and accurate estimation of forest biomass is essential to drive sustainable management of forests. Field-based measurements of forest above-ground biomass (AGB) can be costly and difficult to conduct. Multi-source remote sensing data offers potential to improve the accuracy of modelled AGB predictions. Here, four machine learning methods: Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Classification and Regression Trees (CART) and Minimum Distance (MD) were used to construct forest AGB models of Taiyue Mountain forest, Shanxi Province, China using single and multi-sourced remote sensing data and the Google Earth Engine platform. Results showed that the machine learning method that most accurately predicted AGB was GBDT and spectral index for coniferous (R^2=0.99; RMSE=65.52 Mg/ha), broadleaved (R^2=0.97; RMSE=29.14 Mg/ha), and mixed species (R^2=0.97; RMSE=81.12 Mg/ha) forest types. Models constructed using bivariate variable combinations that included the spectral index improved the AGB estimation accuracy of mixed species (R^2=0.99; RMSE=59.52 Mg/ha) forest types and reduced slightly the accuracy of coniferous (R^2=0.99; RMSE=101.46 Mg/ha), and broadleaved (R^2=0.97; RMSE=37.59 Mg/ha) forest AGB estimation. Overall, parameterising machine learning algorithms with multi-source remote sensing variables can improve the prediction accuracy of mixed species forests.

Keywords: Google Earth Engine; Mixed Species; Landscape; Satellite; Spectral; Waveband

1. Introduction

Remote sensing has great utility in the determination of forest above-ground biomass (AGB) due to the rapid and repeatable acquisition of multi-sensor derived waveband information that correlates with forest biomass structure. Forests cover approximately 40% of the global non-ice land area, and their biomass accounts for about 90% of the terrestrial biomass, as such forests have an irreplaceable role in the terrestrial carbon (C) cycle (Houghton, 2005). Therefore, estimating forest AGB in the study of the C cycle and C stocks in terrestrial ecosystems is of high importance (Vashum, et al. 2012). Traditionally, AGB of forests was determined through manually intensive collection of forest inventory data; however, the emergence of portable terrestrial light detection and ranging (LiDAR) scanners has provided high resolution data to describe forest structure and derive forest inventory metrics (Wulder, et al. 2012), while satellite and airborne LiDAR has enabled the estimation of forest biomass from large areas of inaccessible forest. Developments in remote sensing technology such as synthetic aperture radar (SAR) and interferometric SAR (InSAR) has, particularly through the application of machine learning (ML) techniques, provided further opportunities to improve the accuracy of forest AGB estimates over large areas (Frolking, et al. 2009; Lechner, et al. 2020; Luo, et al. 2020). Recently, LiDAR, optical and radar remote sensing data have been combined into multi-source datasets for research into land cover change and forest biomass estimation (Isbaex, et al. 2021), climate change (He et al. 2023), environmental pollution (Zhang et al. 2023), and forest ecophysiology (Gamon et al. 2023).

The integration of multi-source remote sensing data offers huge potential to improve the predictive power of ML algorithms used in data science. Hyde et al., (2007) demonstrated that the prediction of forest AGB could be improved using a combination of LiDAR, SAR, and InSAR (i.e., LiDAR+SAR/InSAR) rather than using the three types of data individually. Indeed, spatial modelling methods that integrate airborne LiDAR with satellite-based SAR data have been shown to provide spatially explicit AGB estimates over large areas (Tsui et al. 2013). Vafaei et al., (2018) combined multispectral Sentinel-2A imagery with ALOS-2 and PALSAR-2 data to estimate forest AGB using four ML methods. The study revealed that when Sentinel-2A imagery is combined with ALOS-2 and PALSAR-2 data, forest AGB estimates are improved over Sentinel-2A data alone, and that the support vector regression (SVR) method yielded the highest level of accuracy. Similarly, Tamiminia et al., (2022) combined optical, SAR, and airborne LiDAR data to estimate forest AGB using multiple decision tree-based ML methods to reveal that optical and SAR data provided the most accurate estimation of forest AGB; however, there was no significant difference between the ML methods used. Shao et al. (2017) demonstrated the utility of multi-sourced for the AGB estimation of forests by integrating optical (Landsat 8 OLI) and SAR (Sentinel-1A) explanatory variables to parameterize a stacked sparse autoencoder network (SSAE) and show that the data combination outperformed SAR and optical data variables alone for forest AGB estimation over large areas.

Most recently, the accuracy of forest AGB estimation was improved by accounting for tree phenology and dominant tree species with the random forest (RF) method parameterised with LiDAR and Sentinel-1 and Sentinel-2 data (Zhang et al 2023). Consensus in recent literature suggests that radar and optical remote sensing data sources can improve forest AGB estimation over optical or LiDAR data alone (Velasco et al. 2023) and opportunities remain to further refine methodologies by evaluating a broader range of multi-source remote sensing data and ML methods. For example, multi- or hyper-spectral data can be usefully analysed to extract metrics that describe biophysical characteristics of vegetation, and the differences in reflectance spectra of vegetation can also be used to identify specific species at different growth stages (Li, et al. 2012), while transformation of spectral bands using the tassel cap transformation can be used to generate indices that are proxies for texture, frequently used to parameterize models of forest biomass.

Machine learning methods have become prevalent in the development of forest biomass models as they are able to reveal complicated nonlinear relationships in complex datasets. Machine learning methods are widely used because of their adaptiveness, interpretability, and sustainability, and are divided into supervised and unsupervised categories. Supervised learning enables ML algorithms to use training datasets to reveal the relationship between input and output data. Algorithms that require supervised learning include decision trees, logistic regression, support vector machines, and neural networks (Mountrakis, et al. 2011; Rodríguez-Veiga, et al. 2019; Mas, et al. 2008). Whereas unsupervised learning is a data processing method that classifies a large sample of the subject under study through data analysis without category information. Unsupervised classification methods include cluster analysis, principal component analysis and factor analysis (Olaode, et al. 2014). In the ML-based assessment of forest AGB assessment of a single tree species in northern Thailand, the RF method demonstrated higher model accuracy compared to traditional allometric equations and other ML methods (Wongchai, et al. 2022). However, Bulut,

(2023) recommends that multiple ML methods be used with multiple data sources in different environmental conditions to obtain the most accurate forest AGB esitmates.

Commonly used supervised ML methods for forest biomass models include Random Forest (RF; Tian, et al. 2017), Classification and Regression Trees (CART; Breiman, 2017), Gradient Boosting Decision Tree (GBDT; Pham, et al. 2020), and the Minimum Distance (MD; Yang, et al. 2020) method. These ML methods commonly used to estimate forest AGB are evaluated using an R^2 based on the coefficient of determination, root mean square error (RMSE), mean absolute error (MAE), and relative error (RE) (Isbaex, et al. 2021; Han, et al. 2021).

Google Earth Engine (GEE) is a cloud platform that provides powerful tools for processing and analysis of remotely sensed data (Lu, et al. 2016). Through the GEE interface users can access more than 50 petabytes of remote sensing data from Landsat, Sentinel, SAR and digital elevation models (DEM) (Gorelick, et al. 2017). Data processing on the GEE platform can be conducted using Javascript and Python APIs to access Google's compute infrastructure for parallel processing of massive datasets. Recently, scholars have used GEE to analyse environmental change with a focus on forest monitoring (Tamiminia, et al. 2020), conduct large-area multi-source remote sensing-based forest biomass estimation (Yang, et al. 2018) and to develop online visualisation tools (Yan, et al. 2022).

Although several studies have explored the estimation of forest AGB using multi-source remote sensing variables, there is currently no specific construction process to select ML methods and different combinations of remote sensing variables

(Lu, et al. 2006). Here, we use an optimal ML method to construct different forest AGB models using single input datatypes and construct multi-source remote sensing variables for comparison to the optimal single variable. Multi-source remote sensing variable combinations are then constructed according to their importance and correlation between an array of multi-source remote sensing variables to test the optimal forest AGB model. However, to obtain accurate determination of biomass in mixed species forests, it is necessary to consider tree species-specific differences in remotely sensed data. The objectives of this paper are to: (i) improve the estimation of AGB in different forest types i.e., broadleaved, coniferous and mixed species forests; (ii) determine the optimal combination of remotely sensed data to improve the accuracy of forest AGB estimation using ML approaches; and (iii) to explore the forests within the Huodong coal mine area under Taiyue Mountain to validate the selected method.

2. Material and Methods

2.1. Study area

Huodong Mining District (36°30’0’’N, 112°24’0’’E) is a national mining district within Jinzhong coal basin, one of the 14 large coal basins in China delineated in the National Mineral Resources Plan (2016-2020). The mining area is a temperate continental climate, with four distinct seasons and a large temperature difference between day and night. The mean annual temperature is 9.2 °C. The mean annual precipitation is 564.1 mm. It is located in the west of Qinshui Coalfield in Shanxi

Province and covers an area of 4,110 km2 with a total coal resource of 36.6 billion tons. Huodong mining area is not only rich in mineral resources but also has the largest national Forest Park, Taiyue Mountain Forest Park, in Shanxi province. Taiyue Mountain Forest is a species diverse forest with 233 species of woody plants belonging to 44 families and 99 genera: 62 families and more than 500 species of herbaceous plants. The forest total area exceeds 60,000 hectares and is comprised of northern China’s main forest species: Larix principis-rupprechtii, Cunninghamia lanceolata, Pinus tabuliformis with Quercus wutaishanica, Populus spp., Acacia locust, Betula platyphylla. An overview map of the study area is shown in Figure 1.

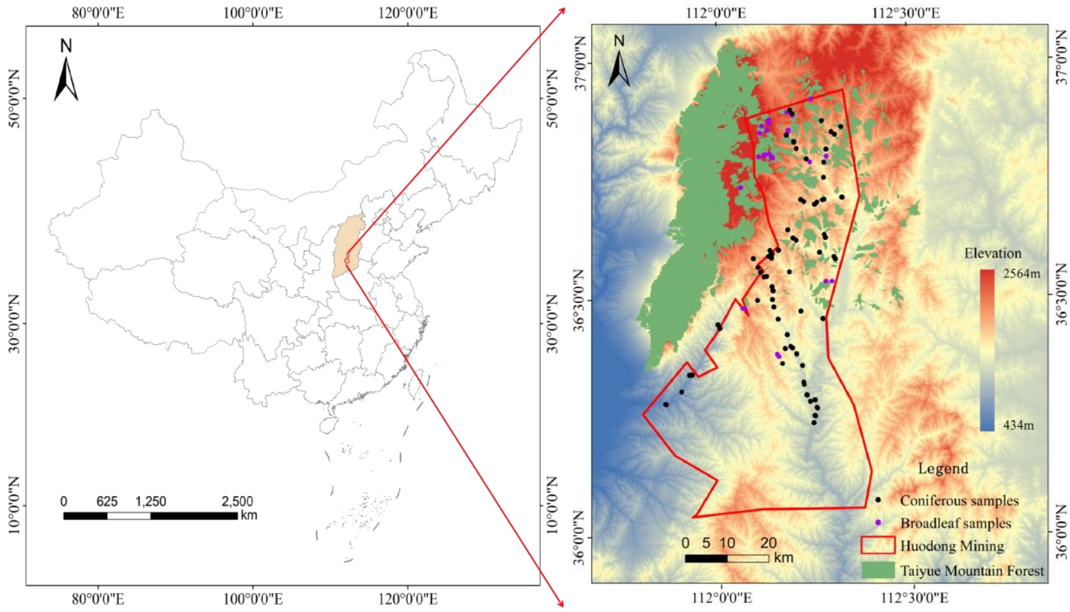

Figure 1. Map of the study area in the southeastern of Shanxi Province, China. The red line box is the Huodong mining area, the green area is the Taiyue Mountain forest, Black and purple points are the coniferous and broadleaved forest sampling sites, respectively.

2.2. Data collection and processing

2.2.1. Data collection

Selection of 128 (30 m × 30 m) forest sample plots (128 mixed, 91 broadleaved, 37 coniferous,) was conducted with the GEE platform using high spatial resolution images to obtain the coordinates of the center point of each forest sample plot (Figure 2). A total of 128 forest sample plots were surveyed between 1st and 23rd August 2022 using a combination of traditional forest mensuration measurements and mobile LiDAR (LiBackpack DGC50) with a relative accuracy of 3 cm, absolute accuracy of 5 cm, scanning frequency of 600,000 points/sec. During the measurement process, the surveyors manually measured diameter at breast height (DBH) and height (H) of all living trees. The ABG of each tree species was estimated using a regional tree species-specific allometric equation (Table 1) (Fang, et al. 2001). All remote sensing data were sourced from datasets available in the GEE cloud platform (https://developers.google.com/earth-engine/datasets/), with the exception of Landsat 8 Level 2, Collection 2, Tier 1 optical data for each 30 m × 30 m forest sample plot Topographic data were obtained from NASA SRTM Digital Elevation, and SAR data were Global PALSAR-2/PALSAR yearly mosaic, and the specific data parameters are shown in Table 2.

Figure 2. Forest sample plot survey (a) single tree diameter at breast height measurement using a diameter at breast height ruler; (b) scanning the forest sample plot using a backpack LiDAR; (c) single tree height measurement using a height gauge; (d) measuring the extent of the forest sample plot using a measuring rope.

Table 1. Allometric equations for estimating the forest species in the study area. AGB is the above-ground biomass (kg), D is the diameter (cm) at breast height (1.3 m), H is the height of the tree (m).

Tree Species	Allometric Equation
Larix principis-rupprechtii	$AGB = 0.2387114(D^2H)^{0.6784}$
Cunninghamia lanceolata	$AGB = 0.00849(D^2H)^{1.10723}$
Populus spp.	$AGB = 0.07363(D^2H)^{0.7745}$
Pinus tabuliformis	$AGB = 0.14187(D^2H)^{0.8728}$
Robinia pseudoacacia	$AGB = 0.02583(D^2H)^{0.6841}$
Quercus wutaishanica	$AGB = 0.04930(D^2H)^{0.8514}$

Table 2. Remote sensing image collection

Name	Earth Engine Snippet	Acquisition Date	Processing Level
Landsat	LANDSAT/LC08/C02/T1_L2	"2022-06-01","2022-08-31"	Level 2

8			
DEM	USGS/SRTMGL1_003	“2000-02-11”	V3
SAR	JAXA/ALOS/PALSAR/YEARLY/SAR	"2020-01-01","2021-01-01"	2.1

2.2.2. Data processing

LiDAR generated 3D cloud point data collected in field for each forest sample plot was preprocessed using the LiDAR360 software (GreenValley International, Zhongguancun Software Park, Haidian). The processes involved forest sample screening and clipping, point cloud data thinning and denoising, ground point cloud segmentation, point cloud normalization and single wood parameter statistics. Finally, the tree height and diameter of single trees in all forest plots were counted separately to obtain the forest biomass of the whole plot.

Processing of the Landsat 8, SAR, and DEM datasets involved filtering to extract the specific study area and removing clouds using a cloud bit mask. Multiple sources of remote sensing variables were selected from specified bands of different image collections to obtain the information shown in Table 3.

2.3. Experimental design

Most of the scientific literature does not explain how to select appropriate variables to develop and evaluate forest AGB models. Based on this knowledge, we designed this experiment to construct forest AGB models using a combination of multi-source remote sensing variables and then compared the accuracy of different variable combinations on forest AGB models to more scientifically follow the optimal combination of single variables and reveal which combination of variables had the best fit.

Four experiments were conducted to assess the utility of different variable combinations and their accuracy in estimating forest AGB: (i) single variable; (ii) multi-source variable combinations; (iii) variable importance; and (iv) Pearson correlation coefficient. The four ML methods (RF, CART, GBDT and MD) used in this study were evaluated with n = 500 decision tree parameters. Each model was analysed by assessing the following four indicators: R^2, RMSE, MAE, and RE. A flowchart that details the satellite-image processing and the generation of forest AGB models using ML is shown in Figure 3.

For model training and validation of the model AGB estimates, the location of each of the 128 forest sample plots was identified using a handheld GNSS receiver (CHC® LT500T, iGage Mapping Corporation, Salt Lake City, USA), and a field-based forest inventory survey conducted.

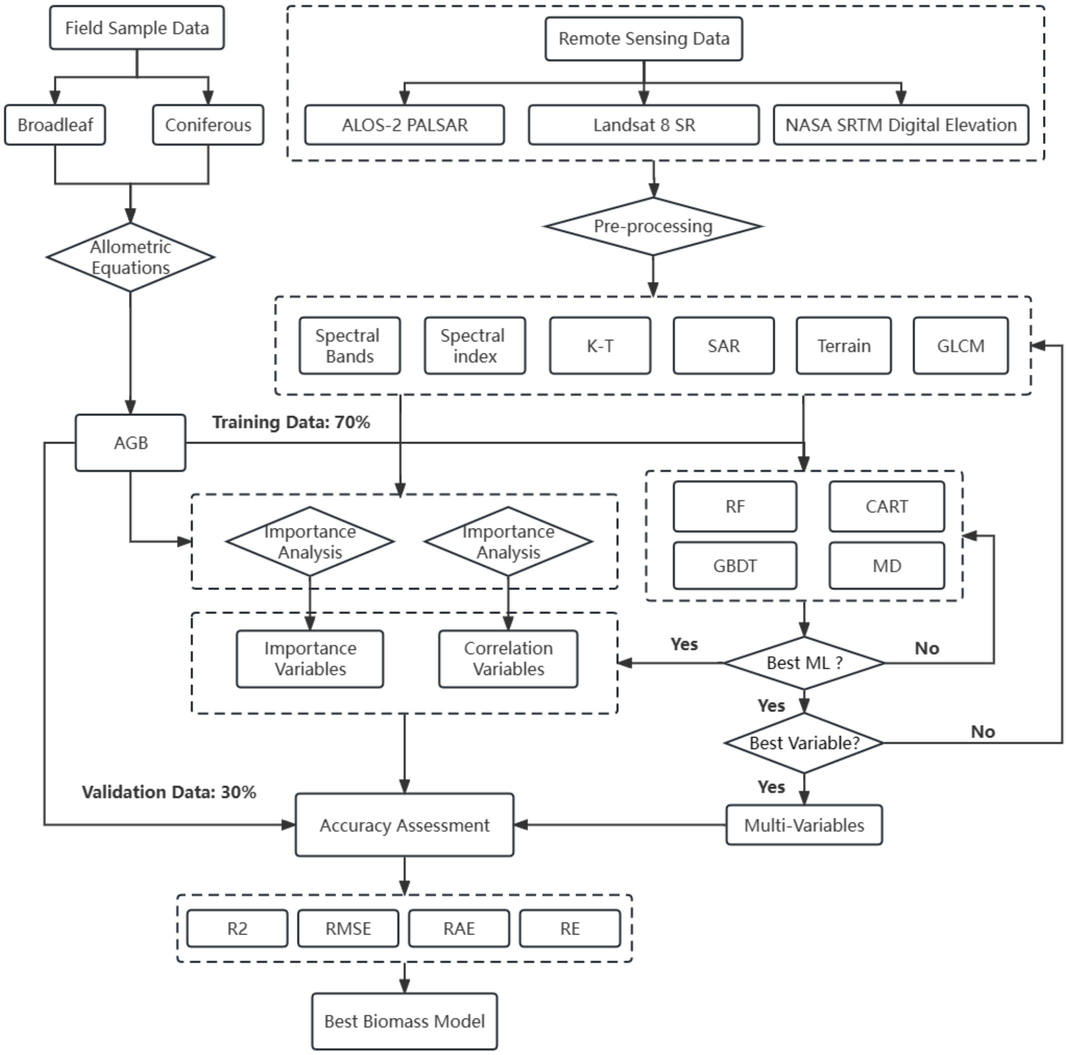

Figure 3. Flowchart for satellite-image processing and the generation of forest above-ground biomass (AGB) models based on machine learning (ML) methods. Among the six variable types obtained during the data processing, the feature variable synthetic aperture radar (SAR) was derived from the ALOS-2 PALSAR data. Spectral bands, spectral indices, Kauth-Thomas (K-T), and gray level co-occurrence matrix (GLCM) all originate from Landsat 8 SR images. Terrain variables were derived from the NASA's shuttle radar topography mission (SRTM).

2.4. Machine learning methods

Four decision tree ML methods (RF, CART, GBDT and MD) were selected from the

ML methods available in the GEE platform to construct a forest biomass model.

2.4.1. Random Forest

Random forest is an integration-based decision tree approach (Cutler, et al. 2012) that is commonly used for classification, regression, and other tasks (Breiman, 2001). It improves the prediction performance by integrating multiple decision trees, each constructed by random subsampling and random feature selection. The random forest approach takes a self-service sampling method (bootstrap sampling), in which k samples are randomly selected from the original dataset to form a collection of subsamples, which can increase the randomness and diversity of the training set and reduce the phenomenon of overfitting. Multiple bootstrap samples are randomly and repeatedly sampled from the training dataset, and then a decision tree is constructed for each bootstrap sample. Finally, the regression results of all decision trees are averaged to obtain the prediction results (Speiser et al. 2019).

2.4.2. Classification and Regression Tree

Classification and regression tree is a decision tree classification and regression method (Loh, et al. 2008). The CART algorithm recursively constructs a decision tree by binary slicing of sample features, where each leaf node represents a decision outcome (Breiman, 2017). For classification problems, the leaf nodes of the decision tree correspond to a category; for regression problems, the leaf nodes of the decision tree correspond to a value. The CART algorithm generates interpretable decision trees with low computational effort and fast training but may produce overfitting for highdimensional data (Gómez, et al. 2012).

2.4.3. Gradient Boosting Decision Tree

The GBDT algorithm is implemented to model and predict data by integrating multiple decision trees where in each iteration step, a decision tree is used to fit the residuals of the current data (Friedman, 2001). Eventually, the predictions of multiple decision trees are weighted and averaged to obtain the final model predictions (Pham et al. 2020). The advantages of the GBDT method are that it can effectively handle many types of data (e.g., numerical, subtypes, and sequential) and can automatically select important features and handle missing data. In addition, the method has a strong generalization capability to handle large-scale datasets and yields good performance in most cases (Li, et al. 2020).

2.4.4. Minimum Distance

The MD method is a classic classification method that classifies samples into different categories by measuring the distance between them (Wolfowitz et al. 1957). The method assumes that there is a variability in distance between the sets of samples of different categories, i.e., the distance between the sets of samples of different categories is farther and the distance between samples of the same category is closer. The basic idea of the shortest distance method is that for a new sample, the distance between it and the sample of each category is calculated, and then it is placed in the category with the closest distance to it (Mahdianpari, et al. 2020). Euclidean distance or Manhattan distance is usually used to measure the distance between samples. In this paper, Euclidean distance is used as a parameter for analysis by default. The advantage of the shortest distance method is its simplicity and ease of use, as well as the fact that it does not require complex prior training and conditioning of the samples. However, it also has some disadvantages, such as sensitivity to outliers and poor performance on unbalanced datasets (Shaharum, et al. 2020).

2.4.5. Model Parameter

The specific parameters of the four machine learning methods used in this paper are shown in Table 3. The RF and GBDT methods have six parameters each, while the CART and MD methods have two parameters each. For a specific parameter explanation, please see the GEE developer documentation available at https://developers.google.com/earth-engine/apidocs

Table 3. Specific parameters of the random forest (RF), classification and regression tree (CART), gradient boosting decision tree (GBDT), minimum distance (MD) machine learning methods

Parameter	RF	CART	GBDT	MD
numberOfTrees	500	-	500	-
variablesPerSplit	14	-	-	-
minLeafPopulation	1	1	-	-
bagFraction	0.5	-	-	-
maxNodes	no limit	no limit	no limit	-
seed	0	-	0	-
shrinkage	-	-	0.005	-
samplingRate	-	-	0.7	-
loss	-	-	LeastAbsoluteDeviation	-
metric	-	-	-	euclidean
kNearest	-	-	-	1

2.5. Biomass model variable

Biomass model variables involved in the construction were divided into six categories, which are the spectral bands of Landsat images, spectral indices, topographic factors, tassel-cap transform (Kauth-Thomas), gray level co-occurrent matrix (GLCM), and SAR factors, where the texture feature variable consist of 18 components, all of which are applied to SR_B2-B7 bands, respectively, and all of the total number of multi-source variables is 156, as shown in Table 4. The specific abbreviated noun explanation is provided in Supplementary.

Table 4. Biomass model single variable

Type of variable	Specific variable factors	number
Landsat bands	Blue(SR_B2), Green(SR_B3), Red(SR_B4), NIR(SR_B5), SWIR1(SR_B6), SWIR2 (SR_B7)	6
Spectral index	NDVI, GNDVI, BNDVI, NDWI, NDWI1, MNDWI, NDMI, NDSI, SIPI, RECl, EVI, EVI2, SR, LAI, GVI, RVI, GRVI, DVI, SAVI, OSAVI, ARVI, VARI, SLAVI, NBR, NDGI, GCVI, GRNDVI, GBNDVI, RBNDVI, RGRI	30
Terrain	Elevation, Slope, Aspect, Hillshade	4
Tassel Cap Transform	Brightness, Greenness, Wetness, TCD(Tasseled Cap Angle), TCA(Tasseled Cap Distance)	5
GLCM	ASM(Angular Second Moment), CONTRAST(Contrast), CORR(Correlation), VAR(Variance), IDM(Inverse Difference Moment), SAVG(Sum Average), SVAR(Sum Variance), SENT(Sum Entropy), ENT(Entropy), DVAR(Difference variance), DENT(Difference entropy), IMCORR1(Information Measure of Corr. 1), IMCORR2(Information Measure of Corr. 2), MAXCORR(Max Corr. Coefficient), DISS(Dissimilarity), INERTIA(Inertia), SHADE(Cluster Shade), PROM(Cluster prominence)	18
SAR	HH(Horizontal transmit/Horizontal receive polarization), HV(Horizontal transmit/Vertical receive polarization)	2

Spectral bands refer to the electromagnetic waves collected at different wavelengths by satellite sensors during the process of acquiring remote sensing images. Different spectral bands have varying reflectivity characteristics for different features. Therefore, extracting different spectral bands in remote sensing images can be utilized to describe and differentiate features. The B2-B7 bands in the Landsat SR data were selected as the spectral band variable.

The spectral index is one of the most important variables for estimation of forest AGB, especially the vegetation index, which is calculated by analysing vegetation reflection or radiation data and can provide information on vegetation growth status, chlorophyll content and vegetation cover (Zeng, et al. 2022). We selected 30 spectral indices as one of the remote sensing variables to participate in the construction of biomass models. Supplementary Table B details all the spectral indices and their abbreviations used in this paper.

Terrain variables play an important role in estimating forest biomass. Elevation can affect the climate and soil conditions, and therefore, the growth of forests. Slope, aspect and hillshade can influence microclimate, rates of soil erosion and water distribution, thereby influencing forest growth and biomass accumulation.

Spectral transformation is the process of converting raw remote sensing image data into a different representation space. Its purpose is to extract the features of different objects in the image for classification, target detection, change detection, and other applications. Spectral transformation is the process of converting raw remote sensing image data to another kind of representation space. In this paper, the five variables in the K-T transformation were selected as the spectral transformation variables.

GLCM is a common texture analysis method based on the second-order combined conditional probability density of the image. It calculates the spatial relationship between different gray levels in the image. Using remote sensing images to monitor forest textural features can capture detailed features within the forest. Through image processing and classification of remote sensing images, various forest types and structural features can be accurately identified. The extraction of the texture features in this paper was performed using the glcmTexture() function in the GEE platform.

The SAR data can be used to estimate the height, density, and volume of vegetation by measuring the radio waves reflected by the vegetation. The HH and HV (polarization backscattering coefficient) bands in the ALOS-2 PALSAR data were selected as the SAR variables for constructing the forest AGB model.

2.6. Model Evaluation

2.6.1. Training and validation datasets

To train and validate the model the 128 plots comprised of coniferous, broadleaved and mixed species forest were allocated into training (70%) and validation (30%) datasets. The number of training and validation of the forest sample points for each tree species are shown in Table 5.

Table 5. Training and validation sample points for different tree species

Forest Type	Training Points	Validation Points	Total
Coniferous	25	12	37
Broadleaved	63	28	91
Mixed	89	39	128

2.6.2. Feature Importance Analysis

Analysis of variable importance using GEE was conducted to determine the magnitude and predictive contribution of optimal variables to the prediction of forest AGB (Zhang, et al. 2019), this analysis method can be used to inform variable selection, model optimization, and interpretation of model prediction results (Li, et al. 2019). Variable importance analysis is a process of determining the importance between all multi-source remote sensing variables and the measured biomass. The biomass of forest sample points is used as training data, and all feature variables as input properties are input into classifiers, such as RF, as classifier attributes. The importance of each feature's relationship with forest AGB was determined using the explain() function in GEE. RF, CART, and GBDT are provided in the developed APP included in this paper (Section 4.3) for variable importance analysis.

2.6.3. Feature correlation

Pearson correlation coefficient (Eq. 1) was used to assess the degree of linear correlation between all multi-source remote sensing variables and the field measurements of forest AGB, which were then ranked from highest to the lowest.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

In the above equation, x_i and y_i are the variables measured, $\bar{x}$ and $\bar{y}$ are the mean values of the predicted and measured, respectively.

2.6.4. Accuracy assessment

The accuracy of each ML model and variable combination was evaluated by validation using data that was not included in the model building process. Four accuracy evaluation indices: coefficient of determination (R^2; Eq. 2), root mean squared error (RMSE, Mg/ha; Eq. 3), mean absolute error (MAE; Eq. 4) and relative error (RE; Eq. 5) were calculated to compare the predicted and observed values. (Cohen, et al. 2009). All of the above evaluation indices were implemented online through the Javascript API of the GEE platform.

$$R^2 = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(a_i - \bar{a})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2(a_i - \bar{a})^2}} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)}{n}} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|p_i - a_i\right| \tag{4}$$

$$RE = \frac{(p_i - a_i)}{a_i} * 100\% \tag{5}$$

In the above expressions, p_i is the forest AGB predicted by the ML model, a_i is the measured mangrove AGB, n is the total number of sampling plots, and $\overline{p}$ and $\overline{a}$ are the mean values of the predicted and measured AGBs, respectively.

3. Results

3.1. Comparison of different methods

The performance of the four ML methods for predicting the coniferous, broadleaved and mixed species forest types using a single spectral variable is shown in Table 6. Irrespective of forest type the R^2 for each of the four ML methods was consistent whilst the differences in RMSE, MAE, and RE metrics enabled selection of the best model. The error metrics of the GBDT method was the smallest and the error metrics of the MD method were largest. Overall, the error metrics of the GBDT method tended to be the smallest and the error metrics of the MD method were the largest. The ranked order of ML method performance by error for broadleaved forest was RF < GBDT < CART < MD, for coniferous forest was GBDT < RF < CART < MD, and for mixed species forest was CART < RF < GBDT < MD. In aggregate, the GBDT method performed best to estimate forest AGB for both univariate and multivariate input datasets.

Table 6. Comparison of random forest (RF), classification and regression tree (CART), gradient boosting decision tree (GBDT), minimum distance (MD) machine learning methods to estimate forest aboveground biomass.

Forest Type	Performance Indicator	Algorithm			
		RF	CART	GBDT	MD
Broadleaved	R^2	0.69	0.69	0.69	0.69

	RMSE	39.59	51.14	40.45	813.97
	MAE	27.80	34.51	28.96	731.24
	RE	0.68	0.83	0.71	18.73
Coniferous	R^2	0.71	0.71	0.71	0.71
	RMSE	80.34	113.40	76.72	646.23
	MAE	61.76	94.31	54.89	576.25
	RE	0.23	0.40	0.20	2.88
Mixed	R^2	0.83	0.83	0.83	0.83
	RMSE	88.67	85.63	89.13	624.61
	MAE	65.89	65.11	66.46	514.01
	RE	0.63	0.47	0.60	8.19

3.2. Single and multi-source variables model evaluation

3.2.1. single variable biomass model construction

The results of the forest AGB models parameterised with a single remotely sensed variable for the three forest types are shown in Table 7. Among the six different univariately constructed models, the RMSE was larger in coniferous forest than in broadleaved and the mixed-species (undifferentiated) forests. For all models with a single variable, spectral index had the highest fit and GLCM had the lowest fit.

For the broadleaved forest type the variable that resulted in the highest correlation between predicted and measured forest AGB was spectral index (R^2=0.97) , however the GLCM variable produced the largest error with an R^2 of 0.01. RMSE, MAE and RE errors of spectral index model are lower than GLCM model. In the coniferous forest spectral index again resulted in the highest R^2 of 0.99, however the GLCM variable produced the lowest correlation with an R^2 of 0.04. Similarly, the strongest correlation of spectral indices in mixed forests had an R2 of 0.97, while the model constructed by GLCM had the lowest accuracy (R2 = 0.02). Consistency with R^2 was demonstrated in the model evaluation results for RMSE and MAE in all forest species.

Table 7. Precision evaluation of single variable models for different tree species.

Forest Type	Variables	R^2	RMSE (Mg/ha)	MAE	RE
Broadleaved	Terrain	0.05	31.00	25.08	0.46
	Band	0.69	40.45	28.96	0.71
	Index	0.97	29.14	21.40	0.35
	SAR	0.17	46.75	33.90	0.61
	K-T	0.10	33.94	28.94	0.52
	GLCM	0.01	30.11	24.45	0.62
Coniferous	Terrain	0.11	74.24	56.64	0.25
	Band	0.71	76.72	54.89	0.20
	Index	0.99	65.52	50.92	0.28
	SAR	0.48	82.68	68.44	0.32
	K-T	0.72	104.93	95.98	0.65
	GLCM	0.04	111.75	93.75	0.95
Mixed	Terrain	0.01	63.99	45.83	0.70
	Band	0.83	89.13	66.46	0.60
	Index	0.97	81.12	51.18	0.61
	SAR	0.22	64.90	47.53	0.82
	K-T	0.48	76.84	52.77	0.55
	GLCM	0.02	92.55	66.92	0.53

The spatial distribution of forest AGB constructed using a single variable for different forest types are shown in Figure 4. The forest AGB of coniferous and broadleaved forests in the region differs greatly, with coniferous forests predominating and broadleaved forests having a more scattered distribution, and the forest AGB of coniferous forest is higher than that of broadleaved forest. The forest biomass distribution without distinguishing tree species (Figure 4b) can more clearly distinguish the difference in forest biomass distribution in the study area.

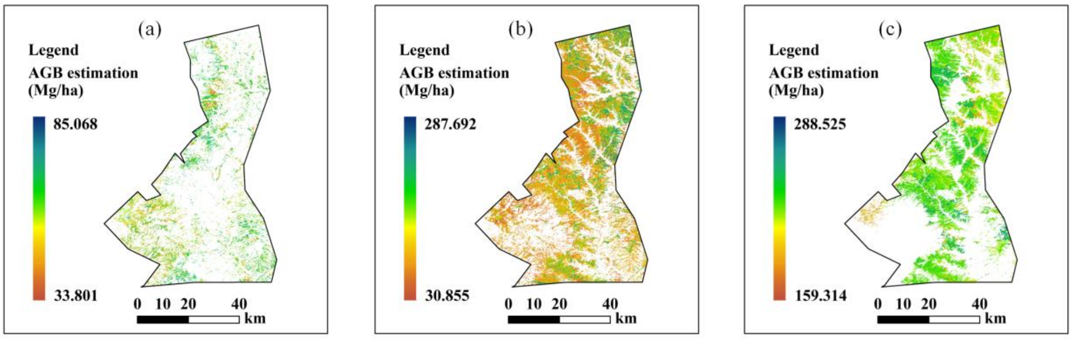

Figure 4. Biomass distribution of single variable (spectral index) for different tree species. (a) broadleaved forest; (b) mixed-species forest; (c) coniferous forest.

3.2.2. Combined biomass model with multi-wavelength variables

In this experiment, 30 variable combinations were compiled Table A (Supplementary), We selected only variable combinations where the model accuracy (R^2) of AGB estimation was > 0.5 as shown in Table 8.

Table 8. Performance comparison of variable combination used in ML to estimate forest AGB.

Variable ID	Variables combination	Mixed		Broadleaved		Coniferous	
		R^2	RMSE	R^2	RMSE	R^2	RMSE
V8	SAR + K-T	0.64	46.94	0.46	34.89	0.43	92.88
V10	Index + Band	0.99	59.52	0.97	37.95	0.99	101.46
V11	Index + K-T	0.98	60.54	0.99	27.68	0.99	109.39
V12	Index + GLCM	0.97	86.41	0.98	49.9	0.99	94.05
V13	Band + K-T	0.91	51.94	0.59	30.61	0.81	85.25
V14	Band + GLCM	0.94	84.21	0.28	28.2	0.82	85.05
V15	K-T + GLCM	0.64	82.21	0.01	38.65	0.46	83.37
V20	Band + Index + K-T	0.58	80.97	0.55	28.18	0.63	73.16
V22	K-T + GLCM + Index	0.65	69.57	0.32	36.6	0.81	96.98
V23	Band + GLCM + Index	0.92	55.24	0.51	26.01	0.7	78.08

Among the combinations of multi-source variables, the highest R^2 (>0.96) between measured and predicted AGB obtained for models using the GBDT method constructed with bivariate combination of spectral indices with spectral bands, K-T transform, and GLCM variables (i.e., V10, V11 and V12, respectively), without distinguishing between forest types.

Models constructed by combining spectral bands with the K-T and GLCM variables had an R^2 >0.8 (mixed-species and coniferous forest types). Based on these results, models were constructed using three, four, and five combinations of variables, but the R^2 values of the models were lower than those of the bivariate models. Among them, the model R^2 values of the coniferous forest type and mixed species (undifferentiated) forest type in the three variable combinations showed consistency in their estimates, but the model fit accuracy of broadleaved forest was much lower. From the V20-V23 multi-source variable combination, it is easy to conclude that coniferous forest outperforms the mixed species forest in terms of fitting accuracy.

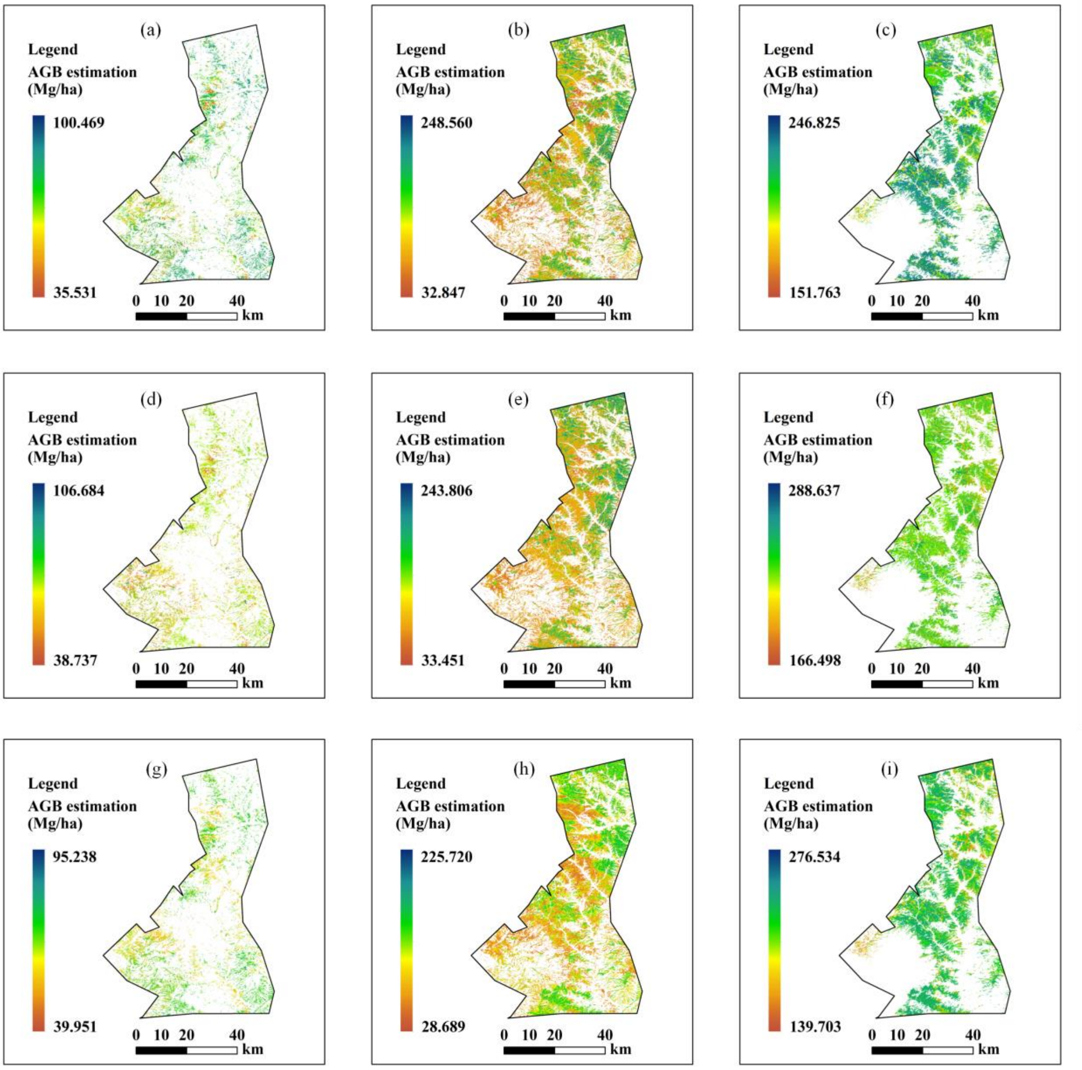

Figure 5. Predicted Forest AGB distribution maps for V10, V11 and V12 variable combinations. Broadleaved forest is shown in panels (a), (d), (f). Mixed species forest shown in panels (b), (e), (h). The coniferous forest shown in (c), (f), (i).

The forest distributions of different forest types have a high degree of consistency (Figure 5). However, because of the differences in the training samples, the predicted values of the forest biomass model are more stable without differentiating the tree species. The coniferous forest biomass predictions had the greatest variation because only 37 sample data point were available, and the coniferous biomass varied more between sample sites.

3.3. Importance analysis of importance variables

Using the importance analysis method in the GEE platform, all the multi-source remote sensing variables were analysed separately with the forest AGB of the sample site, and the importance results were ranked in descending order. Every five variables were stacked in turn to form a new variable combination (Ci_1 to Ci_21) as the input variables of the forest AGB model. Because there were only 105 variables with non-zero values in the results of the variable importance analysis, there were only 21 variable combinations, and the model fitting accuracy results are shown in Figure 6. From the results, it is apparent that there is no strong correlation between the fitting accuracy of variable combinations with varying importance and the number of variables. The biomass models constructed according to the combination of variable importance had low fitting accuracy, and the highest fitting accuracy was only R^2 = 0.23 for Ci_6.

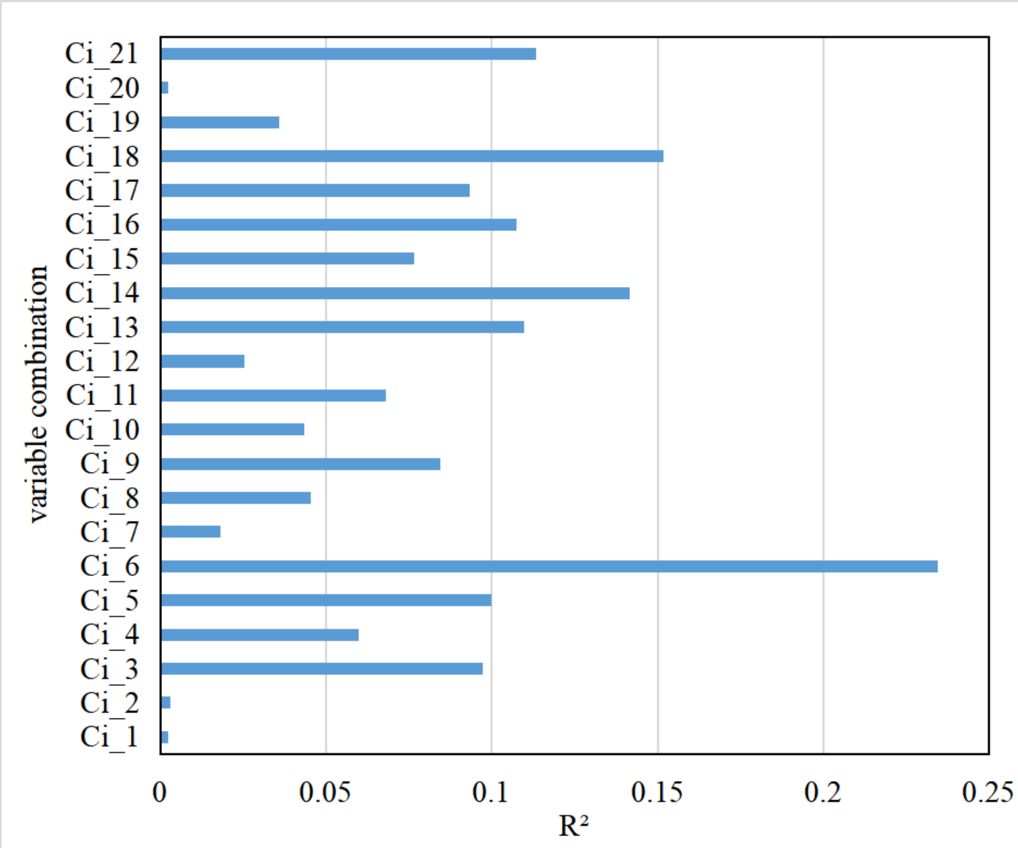

Figure 6. Variable importance model fit R^2 results for 21 different variable combinations.

3.4. Pearson correlation analysis

Pearson correlation analysis was conducted using the RF method for all multi-source remote sensing variables generating a predicted forest AGB and measured forest AGB, and the correlation results were ranked from highest to lowest. First, the five variables with the highest correlation ranking were selected as the initial group of model variables to participate in the biomass model construction. Later, the combination of variables participating in the model construction was added based on the basis of the first group of models, and the cumulative total of 5 variables. Thus, separately validating the resulting forest biomass prediction model after combining the correlation analysis from high to low variables. However, there were only 150 variables with non-empty value values in the results of the variable importance analysis, there were only 30 variable combinations. A new combination of variables (Cp_1 to Cp_30) was formed as the variables of the forest biomass model by superimposing every five variables in turn (Figure 7). In the variable Pearson correlation analysis, the forest biomass model was constructed without distinguishing between tree species in order to reduce the influence of an insufficient number of forest sample points on the results. The results showed that the accuracy of the model fitted by the cumulative equivariant variables tended to first decrease and then gradually increase and stabilise with an increased in the order of correlation of variables. The forest AGB model with the highest accuracy (R^2 = 0.5154) was parameterised using the Cp_22 combination of variables (Figure 7).

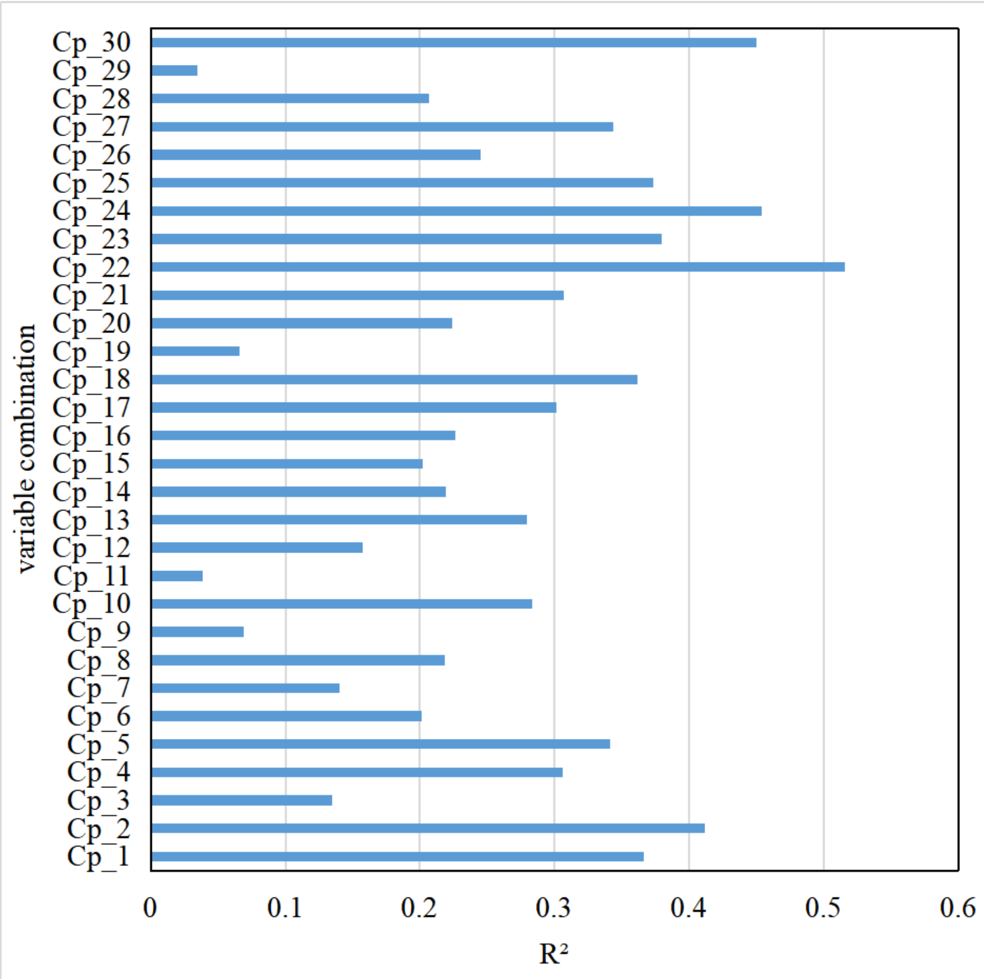

Figure 7. Variable correlation model R^2 results

4. Discussion

The objective of this study was to develop a framework for selecting ML methods and variable combinations to construct a forest AGB model that accurately predicts forest AGB in different forest types. Many studies have reported superior performance of the RF method in predicting forest AGB using remotely sensed data (Chen et al, 2017; Zhang et al, 2023). In this paper, it was found that the GBDT method exhibits higher forest AGB prediction accuracy, particularly when the number of samples points in the training data are large . However, there was not a significant difference between the RF and GBDT methods, which aligns with the findings of previous studies (Tamiminia et al, 2022). The method and process of selecting the optimal forest AGB model used in this study is suitable for all forest AGB modelling. Despite the study area being a mixed species forest located in complex terrain it was still possible to make accurate predictions of forest AGB. By comparing the biomass models built with different variable combinations, the results showed that the number of variables is not directly related to the model accuracy, and in a two-variable combination, the model precision is better than models built with combinations of three or more variables. The forest AGB model built by the variable after importance and correlation screening was less accurate than the optimal single variability combination.

Forest AGB models that do not distinguish between tree species reduce the accuracy of forest AGB estimation. Distinguishing between different tree species to construct species-specific forest AGB models is likely to result in a more accurate assess forest AGB over large areas using remote sensing. However, the construction of species-specific forest AGB models requires a large effort and resource base to obtain forest sample plots for training and validation. In the Huodong coal mine area under Taiyue Mountain forest the broadleaved trees are mostly distributed at lower elevations, leading to the sampling points being located near residential areas and a fragmented distribution of forest sample plots, which may have led to a low overall fit of other single variables with the exception of the spectral index (Zhang et al, 2023). In contrast, coniferous forest was mostly distributed in sparsely populated areas at high altitudes, which makes forest inventory data collection more difficult and explains the limited sample site available for training and validation in this study. Despite the limitations of sample size, it was still possible to estimate coniferous forest AGB with reasonable accuracy because the patches of conifer forests tend to be located in distinct patches that are not often disturbed. However, due to the small sample size available for coniferous species, the construction of variable importance and correlation variables may have led to instability in model fitting accuracy due to insufficient sample points. Therefore, if ML methods are subsequently used for biomass model construction, it is recommended that sufficient sample points be collected to allow for training and validation activities (Qiuli et al, 2023). According to the experimental results of this paper, at least 100 sample points for a single tree species biomass model are needed.

In both univariate and multi-source variable biomass prediction models, the number of samples determines the accuracy of the model, as shown in Figures A and B (Supplementary). Even when no distinction is made between tree species, the prediction of the model for mixed forest AGB were better than those for broadleaved and coniferous forests individually. Among the different combinations of variables, the optimal models that were constructed with spectral indices and K-T best predicted the AGB for broadleaved forests, whilst for coniferous and mixed forests the optimal combination of variables was spectral indices, texture features, spectral indices and bands. In particular, the coniferous forest AGB model parameterised with texture features and spectral indices appeared to compensate for the lower prediction accuracy due to the smaller training and validation sample size.

4.1. Different forest species models

The optimal ML method for estimating forest AGB in the three different forest types was not consistent. Wongchai et al. (2022) reported that many studies have been conducted where different tree species have been analyzed using the same ML methods, with the rationale that canopy information is tree species-specific. In the present study the AGB model prediction error for the three different forest types was ranked from high to low (i.e., coniferous forest > mixed species forest > broadleaved forest) in both single and multi-source variables. The main reason for the higher error in coniferous forest than broadleaved forest is that the sample points collected in broadleaved forest are mostly concentrated near the roadside, where the most abundant tree species is poplar (*Populus* spp.), and the average tree age is similar. However, the sampling data of coniferous forest are concentrated in the higher elevation area, where there is an uneven age distribution, so the difference in sample biomass data is more obvious, which leads to a higher error in broadleaved forest. As the forest inventory plots were sampled in August, all experiments in this paper only considered the prediction and evaluation of forest AGB models during the vegetation growing season, and future model tests will be conducted for different seasons and forest species based on the available results so as to verify the limitations and applicability of the models.

4.2. Accuracy comparison of different combinations of variables

In the ML model construction with a single variable, the optimal forest AGB variable was the spectral index variable, which has been often reported (Wang et al, 2020), followed by the spectral band, irrespective of whether it is a broadleaved, coniferous, or mixed-species forest type with the exception of attempts to parameterise using the GLCM variable. In ML models constructed using multi-source variables, the fitted values based on the spectral index superimposed on other variables were better than the other variable models. The fitted values of the models constructed by equal difference series of variable importance and correlation ranking were lower than those of the single and multi-source models constructed by spectral index variables, regardless of the number. The overall level of model accuracy did not depend on the number of variables, in fact the forest AGB models constructed with single variables with high fit values for multi-source variables provided the most accurate forest AGB estimates. Explanatory variables used in AGB model construction were analysed for multicollinearity using a pairwise comparison of Pearson correlation coefficients, which indicated a strong autocorrelation between the spectral index and the spectral band. Additionally, there was a strong autocorrelation among the SAR HV variables. However, there was no significant autocorrelation observed in the terrain features and GLCM variables. Therefore, incorporating the spectral index/spectral band with other variables can effectively improve the accuracy of the forest AGB model. This is consistent with the results of the multi-source feature variable combinations in Section 3.2.

4.3. Biomass prediction model application

To aid visualisation and interpretation, three GEE-based applications were developed, namely the Forest Biomass and Variable Correlation Analysis Application (https://bqt2000204051.users.earthengine.app/view/forest-agb-variables-correlation-analysis), the Forest Biomass and Variable Importance Analysis Application (https://bqt2000204051.users.earthengine.app/view/forest-agb-variable-importance-analysis) and the Forest Biomass Prediction Application (https://bqt2000204051.users.earthengine.app/view/forest-aboveground-biomass-prediction) to correlate selected multi-source remote sensing variables with the collected forest biomass and to filter the remote sensing variables with high correlation based on correlation coefficients for biomass modelling.

The correlation analysis results for hundreds of variables include correlation coefficients and p-values. The Forest Biomass and Variable Importance Analysis Application performs variable importance analysis based on multi-source remote sensing variables and forest biomass and selects multi-source remote sensing variables for model building based on the variable importance results with the RF, CART, and GBDT ML methods are provided in the variable importance analysis. The Forest Biomass Prediction Application is based on the aforementioned applications but extends them by permitting users to select different ML methods for biomass model prediction using the 30 multi-source variable combinations used in this analysis enabling the assessment of forest AGB estimates and accuracy (i.e., R^2, RMSE, MAE, and RE) to be compared online.

5. Conclusion

In this study, four ML methods were used in the GEE cloud platform to construct forest AGB models using single and multi-source variable combination and their performance evaluated using variable importance values and Pearson correlation coefficients between predicted and measured AGB values. A complete model evaluation system that included R^2, RMSE, MAE, and RE was used to determine best model to predict forest AGB. The results showed the optimal model results were obtained using the GBDT ML method. The most accurate estimation of biomass was achieved for mixed species forests. Multisource remote sensing data and ML methods was able to accurately estimate forest AGB biomass enabling rapid estimation of forest productivity, standing biomass and C stocks in complex topographical landscapes.

Funding

This work was support by the National Key Research and Development Program of China (Intergovernmental and international cooperation in science, technology and innovation) under Grant Number 2022YFE0127700; Royal Society International Exchanges 2022 Cost Share (NSFC) under Grant number IEC\NSFC\223567.

Acknowledgements

The authors sincerely thank the National Aeronautics and Space Administration (NASA) and United States Geological Survey (USGS) for providing the Landsat and DEM data. The authors thank the Japan Aerospace Exploration Agency (JAXA) for providing Global PALSAR-2/PALSAR Yearly Mosaic data. We would like to express our gratitude to Google Earth Engine for offering free cloud computing services. The authors thank the anonymous reviewers for their valuable comments.

Disclosure statement

The authors declare no conflict of interest.

References

Benesty, Jacob, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. "Pearson Correlation Coefficient." *In Noise reduction in speech processing*, 1-4.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45:5-32.

Brovkina, Olga, Jan Novotny, Emil Cienciala, Frantisek Zemek, and Radek Russ. 2017. "Mapping forest aboveground biomass using airborne hyperspectral and LiDAR data in the mountainous conditions of Central Europe." *Ecological Engineering* 100:219-230. https://doi.org/10.1016/j.ecoleng.2016.12.004.

Bulut, Sinan. 2023. "Machine learning prediction of above-ground biomass in pure Calabrian pine (Pinus brutia Ten.) stands of the Mediterranean region, Türkiye." *Ecological Informatics* 74:101951.

Burke, Marshall, and David B Lobell. 2017. "Satellite-based assessment of yield variation and its determinants in smallholder African systems." *Proceedings of the National Academy of Sciences* 114 (9):2189-2194. https://doi.org/10.1073/pnas.1616919114.

Chen, Lin, Chunying Ren, Bai Zhang, Zongming Wang, and Yanbiao Xi. 2018. "Estimation of forest above-ground biomass by geographically weighted regression and machine learning with sentinel imagery." *Forests* 9 (10):582. https://doi:10.3390/f9100582.

Cohen, Israel, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. "Pearson correlation coefficient." Noise reduction in speech processing:1-4.

Fang, Jing-Yun, and Zhang Ming Wang. 2001. "Forest biomass estimation at regional and global levels, with special reference to China's forest biomass." *Ecological Research* 16:587-592.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of statistics*:1189-1232.

Frolking, Stephen, Michael W Palace, DB Clark, Jeffrey Q Chambers, HH Shugart, and George C Hurtt. 2009. "Forest disturbance and recovery: A general review in the context of spaceborne remote sensing of impacts on aboveground biomass and canopy structure." *Journal of Geophysical Research: Biogeosciences* 114 (G2).

Gamon, John A, Ran Wang, and Sabrina E Russo. 2023. "Contrasting photoprotective responses of forest trees revealed using PRI light responses sampled with airborne imaging spectrometry." New Phytologist 238 (3):1318-1332. https://doi.org/10.1111/nph.18754

Gao, Bo-Cai. 1996. "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space." *Remote Sensing of Environment* 58 (3):257-266.

Gitelson, Anatoly A, Yoram J Kaufman, and Mark N Merzlyak. 1996. "Use of a green channel in remote sensing of global vegetation from EOS-MODIS." *Remote Sensing of Environment* 58 (3):289-298.

Gitelson, Anatoly A, Andrés Viña, Verónica Ciganda, Donald C Rundquist, and Timothy J Arkebauer. 2005. "Remote estimation of canopy chlorophyll content in crops." *Geophysical Research Letters* 32 (8). https://doi.org/10.1029/2005gl022688

Gómez, Cristina, Michael A. Wulder, Fernando Montes, and José A. Delgado. 2012. "Modeling Forest Structural Parameters in the Mediterranean Pines of Central Spain using QuickBird-2 Imagery and Classification and Regression Tree Analysis (CART)." *Remote Sensing* 4 (1):135-159. https://doi.org/10.3390/rs4010135.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. "Google Earth Engine: Planetary-scale geospatial analysis for everyone." *Remote Sensing of Environment* 202:18-27. https://doi.org/10.1016/j.rse.2017.06.031.

Hall, Dorothy K, and George A Riggs. 2010. "Normalized-difference snow index (NDSI)." Encyclopedia of snow, ice and glaciers.

Han, Haoshuang, Rongrong Wan, and Bing Li. 2021. "Estimating Forest Aboveground Biomass Using Gaofen-1 Images, Sentinel-1 Images, and Machine Learning Algorithms: A Case Study of the Dabie Mountain Region, China." *Remote Sensing* 14 (1). https://doi.org/10.3390/rs14010176.

He, Kai, Chenjing Fan, Mingchuan Zhong, Fuliang Cao, Guibin Wang, and Lin Cao. 2023. "Evaluation of Habitat Suitability for Asian Elephants in Sipsongpanna under Climate Change by Coupling Multi-Source Remote Sensing Products with MaxEnt Model." *Remote Sensing* 15 (4):1047.

Houghton, R. A. 2005. "Aboveground Forest Biomass and the Global Carbon Balance." *Global Change Biology* 11 (6):945-958. https://doi.org/10.1111/j.1365-2486.2005.00955.x.

Huete, AR, and RD Jackson. 1987. "Suitability of spectral indices for evaluating vegetation characteristics on arid rangelands." *Remote Sensing of Environment* 23 (2):213-IN218.

Huete, Alfredo R. 1988. "A soil-adjusted vegetation index (SAVI)." *Remote Sensing of Environment* 25 (3):295-309.

Hyde, Peter, Ross Nelson, Dan Kimes, and Elissa Levine. 2007. "Exploring LiDAR–RaDAR synergy—predicting aboveground biomass in a southwestern ponderosa pine forest using LiDAR, SAR and InSAR." *Remote Sensing of Environment* 106 (1):28-38. https://doi.org/10.1016/j.rse.2006.07.017.

Jahromi, Mojtaba Naghdyzadegan, Maryam Naghdizadegan Jahromi, Babak Zolghadr-Asli, Hamid Reza Pourghasemi, and Seyed Kazem Alavipanah. 2021. "Google Earth Engine and its application in forest sciences." *Spatial Modeling in Forest Resources Management: Rural Livelihood and Sustainable Development*:629-649.

Jin, Suming, and Steven A. Sader. 2005. "Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances." *Remote Sensing of Environment* 94 (3):364-372. https://doi.org/10.1016/j.rse.2004.10.012.

Jordan, Michael I, and Tom M Mitchell. 2015. "Machine learning: Trends, perspectives, and prospects." *Science* 349 (6245):255-260.

Key, CH, and NC Benson. 1999. "The Normalized Burn Ratio, a Landsat TM radiometric index of burn severity incorporating multi-temporal differencing." *US Geological Survey*:2000.

Lechner, Alex M., Giles M. Foody, and Doreen S. Boyd. 2020. "Applications in Remote Sensing to Forest Ecology and Management." *One Earth* 2 (5):405-412. https://doi.org/10.1016/j.oneear.2020.05.001.

Le Toan, Thuy, S Quegan, MWJ Davidson, Heiko Balzter, Philippe Paillou, Konstantinos Papathanassiou, S Plummer, F Rocca, S Saatchi, and H Shugart. 2011. "The BIOMASS mission: Mapping global forest biomass to better understand the terrestrial carbon cycle." *Remote Sensing of Environment* 115 (11):2850-2860. . https://doi:10.1016/j.rse.2011.03.020.

Li, Deren, Changwei Wang, Yueming Hu, and Shuguang Liu. 2012. "General review on remote sensing-based biomass estimation." *Geomatics and Information, Science of Wuhan University* 37 (6):631-635.

Li, Xiao, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. 2019. "A debiased MDI feature importance measure for random forests." *Advances in Neural Information Processing Systems* 32.

Li, Yingchang, Mingyang Li, Chao Li, and Zhenzhen Liu. 2020. "Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms." *Scientific reports* 10 (1):9952. https://doi: 10.1038/s41598-020-67024-3.

Liu, Hui Qing, and Alfredo Huete. 1995. "A feedback based modification of the NDVI to minimize canopy background and atmospheric noise." *IEEE transactions on geoscience and remote sensing* 33 (2):457-465.

Loh, Wei-Yin. 2008. "Classification and regression tree methods." *Encyclopedia of statistics in quality and reliability* 1:315-323.

Loh, Wei-Yin. 2011. "Classification and regression trees." *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1 (1):14-23.

Lu, Dengsheng. 2007. "The potential and challenge of remote sensing-based biomass estimation." *International Journal of Remote Sensing* 27 (7):1297-1328. https://doi.org/10.1080/01431160500486732.

Lu, Dengsheng, Qi Chen, Guangxing Wang, Lijuan Liu, Guiying Li, and Emilio Moran. 2014. "A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems." *International Journal of Digital Earth* 9 (1):63-105. https://doi.org/10.1080/17538947.2014.990526.

Luo, Weixue, Hyun Seok Kim, Xiuhai Zhao, Daun Ryu, Ilbin Jung, Hyunkook Cho, Nancy Harris, Sayon Ghosh, Chunyu Zhang, and Jingjing Liang. 2020. "New forest biomass carbon stock estimates in Northeast Asia based on multisource data." *Global Change Biology* 26 (12):7045-7066.

Luo, Weixue, Hyun Seok Kim, Xiuhai Zhao, Daun Ryu, Ilbin Jung, Hyunkook Cho, Nancy Harris, Sayon Ghosh, Chunyu Zhang, and Jingjing Liang. 2020. "New forest biomass carbon stock estimates in Northeast Asia based on multisource data." *Global Change Biology* 26 (12):7045-7066.

Lymburner, Leo, Paul J Beggs, and Carol R Jacobson. 2000. "Estimation of canopy-average surface-specific leaf area using Landsat TM data." *Photogrammetric Engineering and Remote Sensing* 66 (2):183-192.

Mahdianpari, M., H. Jafarzadeh, J. E. Granger, F. Mohammadimanesh, B. Brisco, B. Salehi, S. Homayouni, and Q. Weng. 2020. "A large-scale change monitoring of wetlands using time series Landsat imagery on Google Earth Engine: a case study in Newfoundland." *GIScience & Remote Sensing* 57 (8):1102-1124. https://doi.org/10.1080/15481603.2020.1846948.

Mahesh, Batta. 2020. "Machine learning algorithms-a review." *International Journal of Science and Research (IJSR).[Internet]* 9:381-386.

Mas, Jean F, and Juan J Flores. 2008. "The application of artificial neural networks to the analysis of remotely sensed data." *International Journal of Remote Sensing* 29 (3):617-663. https://doi: 10.1080/01431160701352154.

Menze, Bjoern H, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. 2009. "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data." *BMC Bioinformatics* 10:1-16. https://doi.org/10.1186/1471-2105-10-213.

Mountrakis, Giorgos, Jungho Im, and Caesar Ogole. 2011. "Support vector machines in remote sensing: A review." *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (3):247-259. https://doi.org/10.1016/j.isprsjprs.2010.11.001.

Olaode, Abass, Golshah Naghdy, and Catherine Todd. 2014. "Unsupervised classification of images: a review." *International Journal of Image Processing* 8 (5):325-342. https://doi: 10.1016/j.isprsjprs.2010.11.001

Penuelas, J, Frédéric Baret, and I Filella. 1995. "Semi-empirical indices to assess carotenoids/chlorophyll a ratio from leaf spectral reflectance." *Photosynthetica* 31 (2):221-230.

Pham, Tien Dat, Nga Nhu Le, Nam Thang Ha, Luong Viet Nguyen, Junshi Xia, Naoto Yokoya, Tu Trong To, Hong Xuan Trinh, Lap Quoc Kieu, and Wataru Takeuchi. 2020. "Estimating Mangrove Above-Ground Biomass Using Extreme Gradient Boosting Decision Trees Algorithm with Fused Sentinel-2 and ALOS-2 PALSAR-2 Data in Can Gio Biosphere Reserve, Vietnam." *Remote Sensing* 12 (5). https://doi.org/10.3390/rs12050777.

Popescu, Sorin C, Kaiguang Zhao, Amy Neuenschwander, and Chinsu Lin. 2011. "Satellite lidar vs. small footprint airborne lidar: Comparing the accuracy of aboveground biomass estimates and forest structure metrics at footprint level." *Remote Sensing of Environment* 115 (11):2786-2797. https://doi:10.1016/j.rse.2011.01.026.

Rahman, M Mahmudur, and Josaphat Tetuko Sri Sumantyo. 2013. "Retrieval of tropical forest biomass information from ALOS PALSAR data." *Geocarto International* 28 (5):382-403. https://doi:10.1080/10106049.2012.710652.

Rodríguez-Veiga, Pedro, Shaun Quegan, Joao Carreiras, Henrik J. Persson, Johan E. S. Fransson, Agata Hoscilo, Dariusz Ziółkowski, et al. 2019. "Forest biomass retrieval approaches from earth observation in different biomes." *International Journal of Applied Earth Observation and Geoinformation* 77:53-68. https://doi.org/10.1016/j.jag.2018.12.008.

Rondeaux, Geneviève, Michael Steven, and Frédéric Baret. 1996. "Optimization of soil-adjusted vegetation indices." *Remote Sensing of Environment* 55 (2):95-107.

Sazib, Nazmus, Iliana Mladenova, and John Bolten. 2018. "Leveraging the Google Earth Engine for drought assessment using global soil moisture data." *Remote Sensing* 10 (8):1265. https://doi.org/10.3390/rs10081265.

Shaharum, Nur Shafira Nisa, Helmi Zulhaidi Mohd Shafri, Wan Azlina Wan Ab Karim Ghani, Sheila Samsatli, Mohammed Mustafa Abdulrahman Al-Habshi, and Badronnisa Yusuf. 2020. "Oil palm mapping over Peninsular Malaysia using Google Earth Engine and machine learning algorithms." *Remote Sensing Applications: Society and Environment* 17. https://doi.org/10.1016/j.rsase.2020.100287.

Shao, Zhenfeng, Linjing Zhang, and Lei Wang. 2017. "Stacked Sparse Autoencoder Modeling Using the Synergy of Airborne LiDAR and Satellite Optical and SAR Data to Map

Forest Above-Ground Biomass." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10 (12):5569-5582. https://doi.org/10.1109/jstars.2017.2748341.

Sims, Daniel A, and John A Gamon. 2002. "Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages." *Remote Sensing of Environment* 81 (2-3):337-354.

Sinha, Suman, C Jeganathan, LK Sharma, MS Nathawat, Anup K Das, and Shiv Mohan. 2016. "Developing synergy regression models with space-borne ALOS PALSAR and Landsat TM sensors for retrieving tropical forest biomass." *Journal of Earth System Science* 125:725-735. https://doi.org/10.1007/s12040-016-0692-z.

Speiser, Jaime Lynn, Michael E Miller, Janet Tooze, and Edward Ip. 2019. "A comparison of random forest variable selection methods for classification prediction modeling." *Expert Systems with Applications* 134:93-101. https://doi.org/10.1016/j.eswa.2019.05.028.

Su, Yanjun, Qinghua Guo, Baolin Xue, Tianyu Hu, Otto Alvarez, Shengli Tao, and Jingyun Fang. 2016. "Spatial distribution of forest aboveground biomass in China: Estimation through combination of spaceborne lidar, optical imagery, and forest inventory data." *Remote Sensing of Environment* 173:187-199. https://doi.org/10.1016/j.rse.2015.12.002.

Sun, Guoqing, K Jon Ranson, Z Guo, Z Zhang, P Montesano, and D Kimes. 2011. "Forest biomass mapping from lidar and radar synergies." *Remote Sensing of Environment* 115 (11):2906-2916. https://doi.org/10.1016/j.rse.2011.03.021.

Sun, Shaobo, Yafei Wang, Zhaoliang Song, Chu Chen, Yonggen Zhang, Xi Chen, Wei Chen, et al. 2021. "Modelling Aboveground Biomass Carbon Stock of the Bohai Rim Coastal Wetlands by Integrating Remote Sensing, Terrain, and Climate Data." *Remote Sensing* 13 (21). https://doi.org/10.3390/rs13214321.

Tamiminia, Haifa, Bahram Salehi, Masoud Mahdianpari, Lindi Quackenbush, Sarina Adeli, and Brian Brisco. 2020. "Google Earth Engine for geo-big data applications: A meta-analysis and systematic review." *ISPRS Journal of Photogrammetry and Remote Sensing* 164:152-170. https://doi.org/10.1016/j.isprsjprs.2020.04.001.

Tamiminia, Haifa, Bahram Salehi, Masoud Mahdianpari, Colin M Beier, Lucas Johnson, Daniel B Phoenix, and Michael Mahoney. 2022. "Decision tree-based machine learning models for above-ground biomass estimation using multi-source remote sensing data and object-based image analysis." *Geocarto International* 37 (26):12763-12791.https://doi.org/10.1080/10106049.2022.2071475

Tanaka, S, S Goto, M Maki, T Akiyama, Y Muramoto, and K Yoshida. 2007. "Estimation of leaf chlorophyll concentration in winter wheat [Triticum aestivum] before maturing stage by a newly developed vegetation index-RBNDVI." *Journal of the Japanese Agricultural Systems Society (Japan)*.

Tian, Xin, Min Yan, Christiaan van der Tol, Zengyuan Li, Zhongbo Su, Erxue Chen, Xin Li, et al. 2017. "Modeling forest above-ground biomass dynamics using multi-source data and incorporated models: A case study over the qilian mountains." *Agricultural and Forest Meteorology* 246:1-14. https://doi.org/10.1016/j.agrformet.2017.05.026.

Tucker, Compton J. 1979. "Red and photographic infrared linear combinations for monitoring

vegetation." *Remote Sensing of Environment* 8 (2):127-150.
Tsui, Olivier W, Nicholas C Coops, Michael A Wulder, and Peter L Marshall. 2013. "Integrating airborne LiDAR and space-borne radar via multivariate kriging to estimate above-ground biomass." Remote Sensing of Environment 139:340-352. https://doi.org/10.1016/j.rse.2013.08.012.
Vafaei, Sasan, Javad Soosani, Kamran Adeli, Hadi Fadaei, Hamed Naghavi, Tien Dat Pham, and Dieu Tien Bui. 2018. "Improving accuracy estimation of Forest Aboveground Biomass based on incorporation of ALOS-2 PALSAR-2 and Sentinel-2A imagery and machine learning: A case study of the Hyrcanian forest area (Iran)." *Remote Sensing* 10 (2):172.
Vashum, Kuimi T, and S Jayakumar. 2012. "Methods to estimate above-ground biomass and carbon stock in natural forests-a review." *Journal of Ecosystem & Ecography* 2 (4):1-7.
Velasco Pereira, Edward A, María A Varo Martínez, Francisco J Ruiz Gómez, and Rafael M Navarro-Cerrillo. 2023. "Temporal Changes in Mediterranean Pine Forest Biomass Using Synergy Models of ALOS PALSAR-Sentinel 1-Landsat 8 Sensors." *Remote Sensing* 15 (13):3430. https://doi.org/10.3390/rs15133430
Verrelst, Jochem, Michael E Schaepman, Benjamin Koetz, and Matthias Kneubühler. 2008. "Angular sensitivity analysis of vegetation indices derived from CHRIS/PROBA data." *Remote Sensing of Environment* 112 (5):2341-2353. https://doi.org/10.1016/j.rse.2007.11.001.
Wang, Fumin, Huang Jingfeng, Tang Yanlin, and Wang Xiuzhen. 2007. "New vegetation index and its application in estimating leaf area index of rice." *Rice Science* 14 (3):195-203.
Wang, Dezhi, Bo Wan, Jing Liu, Yanjun Su, Qinghua Guo, Penghua Qiu, and Xincai Wu. 2020. "Estimating aboveground biomass of the mangrove forests on northeast Hainan Island in China using an upscaling method from field plots, UAV-LiDAR data and Sentinel-2 imagery." *International Journal of Applied Earth Observation and Geoinformation* 85:101986.
Willmott, Cort J, Steven G Ackleson, Robert E Davis, Johannes J Feddema, Katherine M Klink, David R Legates, James O'donnell, and Clinton M Rowe. 1985. "Statistics for the evaluation and comparison of models." *Journal of Geophysical Research: Oceans* 90 (C5):8995-9005.
Wolfowitz, Jacob. 1957. "The minimum distance method." *The Annals of Mathematical Statistics*:75-88.
Wongchai, Warakhom, Thossaporn Onsree, Natthida Sukkam, Anucha Promwungkwa, and Nakorn Tippayawong. 2022. "Machine learning models for estimating above ground biomass of fast growing trees." *Expert Systems with Applications* 199:117186.
Wulder, Michael A, Joanne C White, Ross F Nelson, Erik Næsset, Hans Ole Ørka, Nicholas C Coops, Thomas Hilker, Christopher W Bater, and Terje Gobakken. 2012. "Lidar sampling for large-area forest characterization: A review." *Remote Sensing of Environment* 121:196-209.
Xu, Hanqiu. 2007. "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery." International Journal of Remote

Sensing 27 (14):3025-3033. https://doi.org/10.1080/01431160600589179.

Yan, Xingguang, Jing Li, Di Yang, Jiwei Li, Tianyue Ma, Yiting Su, Jiahao Shao, and Rui Zhang. 2022. "A Random Forest Algorithm for Landsat Image Chromatic Aberration Restoration Based on GEE Cloud Platform—A Case Study of Yucatán Peninsula, Mexico." *Remote Sensing* 14 (20). https://doi.org/10.3390/rs14205154.

Yang, Lu, Shunlin Liang, and Yuzhen Zhang. 2020. "A New Method for Generating a Global Forest Aboveground Biomass Map From Multiple High-Level Satellite Products and Ancillary Information." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13:2587-2597. https://doi.org/10.1109/jstars.2020.2987951.

Yang, Qiuli, Chunyue Niu, Xiaoqiang Liu, Yuhao Feng, Qin Ma, Xuejing Wang, Hao Tang, and Qinghua Guo. 2023. "Mapping high-resolution forest aboveground biomass of China using multisource remote sensing data." *GIScience & Remote Sensing* 60 (1):2203303. https://doi.org/10.1080/15481603.2023.2203303

Yang, Wei, Hideki Kobayashi, Cong Wang, Miaogen Shen, Jin Chen, Bunkei Matsushita, Yanhong Tang, et al. 2019. "A semi-analytical snow-free vegetation index for improving estimation of plant phenology in tundra and grassland ecosystems." *Remote Sensing of Environment* 228:31-44. https://doi.org/10.1016/j.rse.2019.03.028.

Yang, Zelong, Wenwen Li, Qi Chen, Sheng Wu, Shanjun Liu, and Jianya Gong. 2018. "A scalable cyberinfrastructure and cloud computing platform for forest aboveground biomass estimation based on the Google Earth Engine." *International Journal of Digital Earth* 12 (9):995-1012. https://doi.org/10.1080/17538947.2018.1494761.

Zhang, Linjing, Xiaoxue Zhang, Zhenfeng Shao, Wenhao Jiang, and Huimin Gao. 2023. "Integrating Sentinel-1 and 2 with LiDAR data to estimate aboveground biomass of subtropical forests in northeast Guangdong, China." *International Journal of Digital Earth* 16 (1):158-182. https://doi.org/10.1080/17538947.2023.2165180

Zhang, Xiang, Lexin Li, Hua Zhou, Yeqing Zhou, and Dinggang Shen. 2019. "Tensor generalized estimating equations for longitudinal imaging analysis." *Statistica Sinica* 29 (4):1977.

Zhang, Yuzhen, Jun Ma, Shunlin Liang, Xisheng Li, and Manyao Li. 2020. "An Evaluation of Eight Machine Learning Regression Algorithms for Forest Aboveground Biomass Estimation from Multiple Satellite Data Products." *Remote Sensing* 12 (24). https://doi.org/10.3390/rs12244015.

Zhang, Yali, Ni Wang, Yuliang Wang, and Mingshi Li. 2023. "A new strategy for improving the accuracy of forest aboveground biomass estimates in an alpine region based on multi-source remote sensing." *GIScience & Remote Sensing* 60 (1):2163574. https://doi.org/10.1080/15481603.2022.2163574

Zhao, Yifan, Weiwei Zhu, Panpan Wei, Peng Fang, Xiwang Zhang, Nana Yan, Wenjun Liu, Hao Zhao, and Qirui Wu. 2022. "Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period." *Ecological Indicators* 135. https://doi.org/10.1016/j.ecolind.2021.108529.

Zhang, Zheyuan, Jia Wang, Nina Xiong, Boyi Liang, and Zong Wang. 2023. "Air Pollution Exposure Based on Nighttime Light Remote Sensing and Multi-source Geographic Data in Beijing." *Chinese Geographical Science* 33 (2):320-332.