Integrating Pigeon-Inspired Optimization and Support Vector Machines for Forest Aboveground Biomass Estimation

Xiaomeng Kang , Ling Wang * , Chunyan Chang , Xicun Zhu , Xiao Liu , Chang Qiu , Xianzhang Meng , Danning Chen

Posted Date: 27 March 2026

doi: 10.20944/preprints202603.2261.v1

Keywords: aboveground biomass; Sentinel-2; Pigeon-inspired optimization; support vector machine; forest ecosystems

 doi:10.20944/preprints202603.2261.v1

Article

Integrating Pigeon-inspired optimization and Support Vector Machines for Forest Aboveground Biomass Estimation

Xiaomeng Kang, Ling Wang *, Chunyan Chang, Xicun Zhu, Xiao Liu, Chang Qiu, Xianzhang Meng and Danning Chen

College of Resources and Environment, Shandong Agricultural University, Tai'an 271018, China

* Correspondence: lingwang@sdau.edu.cn

Abstract

Accurate estimation of aboveground biomass (AGB) in mountainous forest ecosystem remains a significant challenge due to complex terrain, the high cost and limited applicability of traditional field-based methods. To address this issue, a remote sensing–based AGB estimation framework integrating intelligent optimization and machine learning was developed for Mount Tai in eastern China. Sentinel-2 multispectral data were selected to derive 105 candidate variables, including spectral bands, vegetation indices, texture features, and topographic factors, from which 17 key variables were selected using Pearson correlation analysis for model construction. A Support Vector Machine (SVM) optimized by the Pigeon-inspired optimization (PIO) algorithm was developed to adaptively determine optimal hyperparameters, and its performance was compared with that of Random Forest (RF) and standard SVM models. The results demonstrate that the PIO-SVM model achieved the best overall performance. For the training dataset, the model obtained an R^2 of 0.85. For the test dataset, the R^2 reached 0.73, outperforming RF (0.70) and standard SVM (0.72). The spatial distribution of AGB derived from the optimal model shows higher AGB values in the central and northern regions characterized by dense forest cover, in close agreement with field observations. These findings indicate that the PIO algorithm effectively enhances SVM hyperparameter optimization in complex parameter spaces, significantly improving the accuracy and stability of AGB estimation in mountainous forest. This study provides a reliable and efficient framework for regional-scale monitoring of forest biomass and carbon sink dynamics.

Keywords: aboveground biomass; Sentinel-2; Pigeon-inspired optimization; support vector machine; forest ecosystems

1. Introduction

Forest ecosystems constitute a fundamental component of the global terrestrial carbon cycle and play an indispensable role in maintaining ecological security, regulating climate, and enhancing carbon sequestration. As a key indicator for evaluating carbon storage and ecosystem functioning, AGB serves as a critical metric for quantifying carbon sink capacity and analyzing carbon cycling processes [1]. Accurate characterization of the spatial distribution of AGB is therefore essential for providing scientific support for carbon peaking and carbon neutrality strategies, as well as for guiding sustainable regional ecosystem management [2,3].

Conventional AGB estimation methods primarily rely on field plot surveys combined with allometric growth equations. Although these approaches may achieve high accuracy at local scales, they are constrained by substantial labor and financial costs, ecological disturbance, and limited spatial representativeness. Consequently, the large-scale and efficient utilization of them, particularly in complex mountainous areas, still presents challenges [4,5]. With the rapid development of remote sensing technologies, non-destructive AGB estimation methods that integrate multi-source satellite

 doi:10.20944/preprints202603.2261.v1

imagery with ground-based observations have emerged as a dominant research paradigm [6,7]. However, the relationships between remote sensing variables and forest structural parameters are often highly nonlinear and scale-dependent, which limits the ability of traditional linear regression models to capture these complex interactions [8]. In this context, machine learning approaches, such as random forests and support vector machines, have emerged as key techniques for forest biomass estimation from remote sensing data due to their strong capability to model high-dimensional and nonlinear relationships.

Machine learning models have been widely applied in forest aboveground biomass estimation. For example, Foody et al. compared multiple linear regression and backpropagation (BP) neural networks for tropical forest biomass estimation using Landsat TM data [9]. Zhang et al. developed a SVM model based on Landsat-7 spectral variables and achieved relatively high estimation accuracy [10]. Dang et al. proposed a robust forest biomass estimation model using Sentinel-2 imagery by extracting a diverse set of feature variables [11]. In China, Wang Yifu et al. [12] and Lin Zhuo et al. [13] developed biomass estimation models for specific tree species using BP neural networks and SVMs, respectively, while Zhang Pengchao et al. employed MODIS data in combination with a random forest model to estimate forest AGB across the Qinghai–Tibet Plateau [14]. Collectively, these studies demonstrate the effectiveness of machine learning approaches for forest biomass estimation, while also indicating that model performance is influenced by multiple factors, including forest type, terrain complexity, and feature selection. To address the reliance of traditional machine learning models on manual expertise for hyperparameter tuning and to further improve model generalization and stability in complex environments, this study introduces and applies a swarm intelligence–based optimization method, namely the PIO algorithm. Originally proposed by Duan Haibin in 2014, this algorithm simulates pigeons' homing behavior through geomagnetic navigation and landmark recognition, thereby achieving a balance between global exploration and local exploitation. As a result, it demonstrates advantages such as fast convergence and strong global optimization capability [15]. Previous studies suggest that integrating the PIO algorithm with machine learning models, particularly SVMs, can improve prediction accuracy and robustness; however, its systematic application to forest AGB estimation using remote sensing data in complex mountainous regions, as well as comprehensive comparisons with mainstream models, remains limited.

Mount Tai, a representative temperate forest ecosystem in eastern China, provides an ideal study area for remote sensing–based estimation of forest parameters in complex terrain due to its pronounced topographic variability and diverse vegetation, which is dominated primarily by pine–oak forests. Focusing on the Mount Tai region in Shandong Province, this study integrates Sentinel - 2 multispectral imagery with ground - based plot survey data to develop and systematically compare the performance of RF, SVM and PIO-SVM models for AGB estimation. The principal innovation of this research lies in the first application of the PIO algorithm to forest AGB estimation in Mount Tai's mountainous environment. By automatically optimizing key SVM parameters, this approach aims to overcome the limitations of conventional parameter tuning methods and enhance model accuracy and robustness. Through comparative analysis, this study seeks to identify the most suitable remote sensing–based AGB estimation model for the region, thereby providing methodological guidance and scientific support for precise forest carbon sink monitoring, ecosystem conservation, and management in Mount Tai and other mountainous forest ecosystems.

2. Materials and Methods

2.1. Overview of the Study Area

The study area is situated in Taian City, Shandong Province, China, encompassing the Mount Tai region (116°50′–117°12′E, 36°11′–36°31′N). Located at the boundary between Jinan and Taian, the area covers approximately 426 km² and includes the highest peak on the North China Plain. The region is characterized by a warm temperate continental monsoon climate, with a long-term mean annual temperature of 12.6 °C and a mean annual precipitation of approximately 758 mm, most of

 doi:10.20944/preprints202603.2261.v1

which occurs between June and September. The terrain is characterized by steep mountainous landscapes, with elevation gradually decreasing from northwest to southeast. The highest point, Jade Emperor Peak, reaches an elevation of 1,545 m above sea level, and mountainous terrain accounts for more than 90% of the total area. Forests cover approximately 80% of the region, and the dominant vegetation types include deciduous broadleaf forests (e.g., *oak species*), coniferous forests (e.g., *Pinus tabulaeformis* and *Platycladus orientalis*), and mixed forests. In addition, portions of the study area fall within the protection zone of a World Cultural and Natural Heritage site.

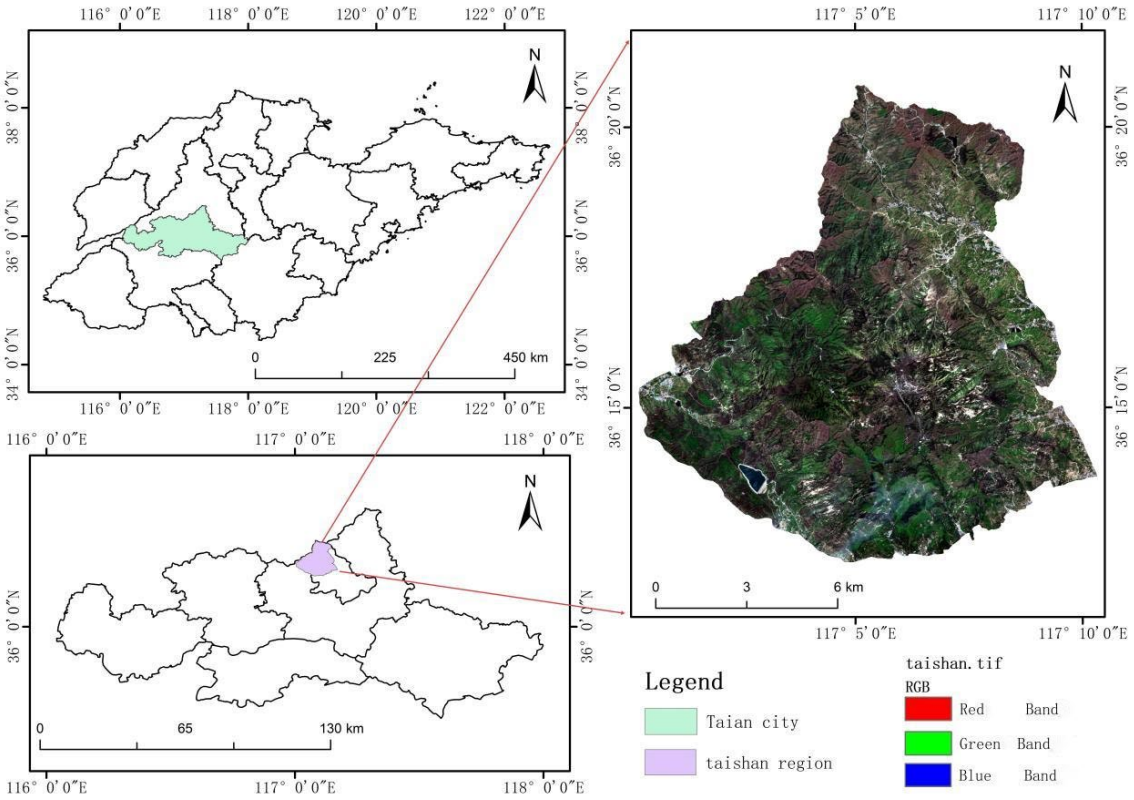


Figure 1. Location of the study area.

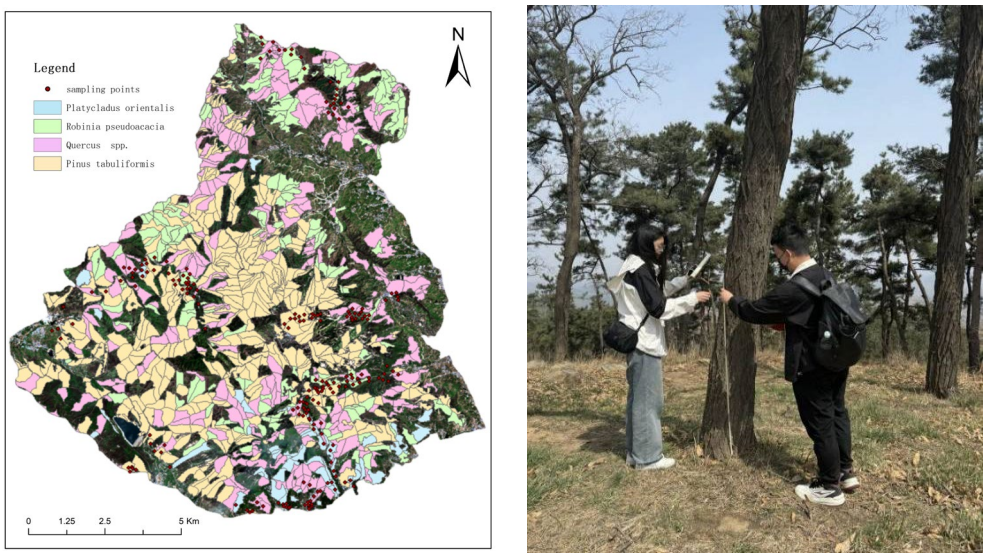
2.2. Remote Sensing Data

The Sentinel-2 remote sensing imagery used in this study was acquired from the European Space Agency (ESA) Copernicus Data Centre on 27 June 2024. Sentinel-2A scenes with cloud cover of less than 5% were selected to minimize atmospheric interference. Image pre - processing, encompassing resampling, radiometric calibration, atmospheric correction, and spatial subsetting, was carried out via SNAP and ENVI software to generate surface reflectance products. Sentinel-2 is a core component of the ESA Copernicus program and consists of two satellites, Sentinel-2A and Sentinel-2B. Its primary mission is to provide high-resolution multispectral imagery of global land surfaces, coastal zones, and inland waters, thereby supporting applications such as agricultural monitoring, forest management, land-use change detection, disaster assessment, and environmental monitoring. The onboard Multispectral Instrument (MSI) acquires multispectral data in 13 spectral bands spanning the visible, near-infrared (VNIR), and shortwave infrared (SWIR) regions, with spatial resolutions of 10 m (four bands), 20 m (six bands), and 60 m (three bands). The spectral configuration of Sentinel-2 is optimized for vegetation monitoring. In particular, the red-edge bands (Bands 5, 6, and 7) enable detailed assessment of chlorophyll content and vegetation condition, while the high-resolution visible bands (Bands 2, 3, and 4) support accurate land-cover and land-use classification. Compared with similar satellite platforms, Sentinel-2 offers improved data accessibility and revisit frequency, providing essential data support for global environmental and natural resource research.

 doi:10.20944/preprints202603.2261.v1

2.3. Ground Survey Data

Field data were collected between May and June 2024. The dominant tree species—including *Quercus spp., Robinia pseudoacacia, Pinus tabuliformis, and Platycladus orientalis*—were identified through field investigations combined with existing forestry inventory data. Following the principles of species representativeness, uniform spatial distribution, and measurement feasibility, representative sample plots with a size of 10 m × 10 m were established for each species. The geographic coordinates and elevation of the center of each plot were recorded using a GPS receiver, and detailed stand attributes within each plot were measured. In total, 205 sample plots were surveyed across the four target species (*Pinus tabuliformis, Platycladus orientalis, Robinia pseudoacacia, and Quercus spp.*). The collected field data included diameter at breast height (DBH), root collar diameter, tree height, and crown width. Due to constraints such as steep mountainous terrain and forest fire prevention regulations, the sample plots could not be uniformly distributed across the entire study area. The spatial distribution of the sample plots is illustrated in Figure 2.



(**a**) (**b**)

Figure 2. (**a**) Distribution of sample sitess and pecies distribution; (**b**) data collection.

The allometric growth equations used for calculate aboveground biomass in this study were those recommended in the *Guidelines on Carbon Accounting and Monitoring for Afforestation Project* (National Forestry and Grassland Administration of China; https://www.forestry.gov.cn/). Regional AGB was calculated based on field sampling data collected within the study area, the mean values appeared in Table 1 and the specific equations employed are listed in Table 2. In these equations, diameter at breast height (DBH) is expressed in centimeters (cm), tree height is measured in meters (m), and biomass is reported in metric tons per hectare (t/hm²). For tree species lacking region-specific allometric equations, biomass estimates were obtained using equations developed for ecologically similar neighboring regions.

Table 1. Summary of Field Survey Data.

Tree Specias	Number	Minimum	Maximum	Averages	Standard Deviation	Coefficient of Variation (%)
Platycladus orientalis	45	70.4047	519.5040	260.6584	123.9561	47.5550
Quercus spp	47	50.1650	541.6702	202.2876	127.4144	62.9867
Pinus tabuliformis	53	45.9000	567.1940	241.6070	133.9167	55.4275
Robinia pseudoacacia	60	43.0483	460.9291	222.9042	109.0712	48.9319

 doi:10.20944/preprints202603.2261.v1

Table 2. Allometric Growth Model of Major Tree Species in the Study Area.

seeds of trees	Allometric Growth Equation	References
Platycladus orientalis	1.034885+0.0223（D2H）+0.095（D2H）0.571+0.714（D2H）0.583	[17]
Quercus spp	0.0369（D2H）0.9165+0.00051（D2H）1.3377+0.00021（D2H）1.171	[17]
Pinus tabuliformis	0.946835+0.214（D2H）+0.77（D2H）0.679+0.848（D2H）0.594	[17]
Robinia pseudoacacia	0.05527（D2H）0.8576+0.02425（D2H）0.7908+0.0545（D2H）0.4574	[17]

2.4. Methodological Flowchart

The overall workflow of this study involves the extraction of a comprehensive set of remote sensing variables related to forest aboveground biomass using ARCGIS software, followed by their integration with AGB estimates derived from field sampling data. After performing correlation analysis and feature selection, multiple machine learning models are developed and evaluated in terms of estimation accuracy. This process ultimately results in the generation of a spatial distribution map of forest AGB across the study area, as illustrated in Figure 3.

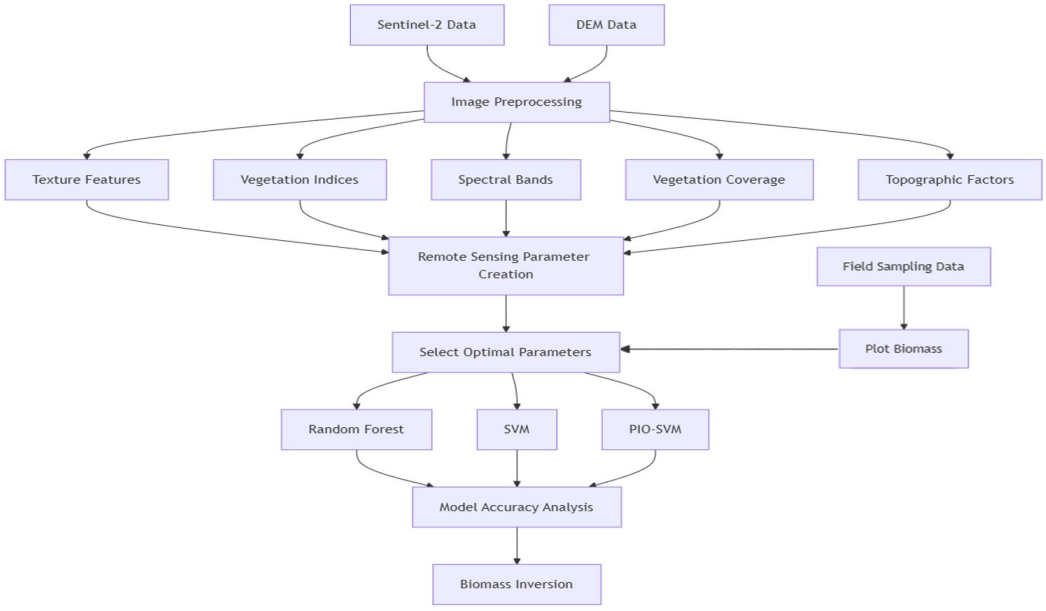


Figure 3. Method flow chart.

2.5. Feature Variable Extraction and Selection

In forest AGB estimation studies, spectral variables and textural features are widely used as predictor variables. Among these, vegetation indices effectively characterize vegetation growth conditions and canopy cover dynamics and have therefore been extensively applied in vegetation monitoring and AGB estimation [17]. In this study, Sentinel-2 multispectral imagery was utilized, including Bands 2 (blue), 3 (green), 4 (red), 5–7 (red-edge), 8 (near-infrared), and 11–12 (shortwave infrared).. Based on these spectral bands, six widely used vegetation indices were derived: the Normalized Difference Vegetation Index (NDVI), Fractional Vegetation Cover (FVC), Enhanced Vegetation Index (EVI), Normalized Difference Red Edge Index (NDRE), Optimized Soil-Adjusted Vegetation Index (OSAVI), and Kernel Normalized Difference Vegetation Index (KNDVI).

 doi:10.20944/preprints202603.2261.v1

Texture features were additionally incorporated to characterize the spatial structure and heterogeneity of forest vegetation, thereby improving AGB estimation accuracy [18]. To reduce information redundancy and extract representative texture information [19], texture metrics were derived from the selected 10-band imagery, namely mean, variance, homogeneity, contrast, dissimilarity, and entropy. Window size selection is critical in texture analysis: overly small windows are prone to noise, whereas excessively large windows may obscure spatial heterogeneity. Accordingly, texture features were extracted using the Gray-Level Co-occurrence Matrix (GLCM) method with a 3×3 moving window, which provides a balance between detail preservation and feature representativeness. Principal Component Analysis (PCA) was subsequently applied to reduce feature dimensionality while retaining at least 90% of the total variance [21].

In addition, three components derived from the Tasseled Cap Transformation—brightness, greenness, and wetness—were combined with topographic variables to construct a comprehensive feature set for AGB estimation. Terrain factors were derived from a 30 m resolution Digital Elevation Model (DEM) acquired from the Geospatial Data Cloud platform, including slope, aspect, terrain wetness index (TWI), and terrain solar radiation index. The incorporation of these variables improves the model's ability to account for complex terrain effects and ecological variability.

Table 3. lists the feature variables used in this study.

Variables	Feature Variable	Formula	References
Spectral Variable	Band reflectance (Sentinel-2 bands)	Band i (i = 2,3, …, 12)	[22]
	Normalized difference vegetation index (NDVI)	(NIR − R) / (NIR + R)	[23]
	Enhanced vegetation index (EVI)	2.5 × (NIR − R) / (NIR + 6R − 7.5B + 1)	[24]
	Fractional Vegetation Cover(FVC)	(NDVI − NDVI_min) / (NDVI_max − NDVI_min)	[25]
	Normalized Difference Red Edge Index(NDRE)	(NIR - RE) / (NIR + RE)	[26]
	Optimized Soil Adjusted Vegetation Index(OSAVI)	(1+0.16)(NIR - R) / (NIR + R + 0.16)	[27]
	Kernel Normalized Difference Vegetation Index(KNDVI)	Tanh ($NDVI^2$)	[28]
Texture feature	Mean	$\sum_{i,j=0}^{N-1} ip_{ij}$	[29]
	Contrast	$\sum_{i,j=0}^{N-1} ip_{ij}(i-j)^2$	[29]
	Variance	$\sum_{i,j=0}^{N-1} p_{ij}(i=ME)^2$	[29]
	Dissimilarity	$\sum_{i,j=0}^{N-1} ip_{ij}\lvert i-j\rvert$	[29]
	Homogeneity	$\sum_{i,j=0}^{N-1} i\frac{p_{ij}}{1+(i-j)^2}$	[29]
	Entropy	$\sum_{i,j=0}^{N-1} ip_{ij}(-\ln P_{ij})$	[29]
	Correlation	$\sum_{i,j=0}^{N-1} iPij\left[\frac{(i-ME)(j-ME)}{\sqrt{VA_iVA_i}}\right]$	[29]
	Second moment	$\sum_{i,j=0}^{N-1} ip_{ij}^2$	[29]
Tasseled Cap Transformation	Brightness	0.3510B + 0.3813G + 0.3437R + 0.7196NIR + 0.2396SWIR1 + 0.1949SWIR2	[30]

 doi:10.20944/preprints202603.2261.v1

Variables	Feature Variable	Formula	References
	Greenness	−0.3599B − 0.3533G − 0.4734R + 0.6633NIR + 0.0087SWIR1 − 0.2856SWIR2	[30]
	Wetness	0.2578B + 0.2305G + 0.0883R + 0.1071NIR − 0.7611SWIR1 − 0.5308SWIR2	[30]

Notes: B, G, and R denote the blue, green, and red bands, respectively; NIR denotes the near-infrared band; RE denotes the red-edge band; SWIR1 and SWIR2 represent shortwave infrared bands 1 and 2, respectively.

2.6. Pearson Correlation

Before constructing the estimation models, Pearson correlation analysis was performed to identify feature variables that are significantly associated with forest AGB, thereby reducing feature redundancy and mitigating the risk of model overfitting. This analysis is used to quantify the strength of linear relationships between candidate independent variables and AGB. The Pearson correlation coefficient r is defined as:

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

where n is the sample size; x_i and y_i represent the observed values of variables x and y at sample point i, respectively; and $\bar{x}$ and $\bar{y}$ denote the mean values of variables x and y. The correlation coefficient r ranges from −1 to 1, with its absolute value indicating the strength of linear association between the variables. In this study, the r value was jointly considered with the corresponding p-value obtained from statistical significance testing, with $p < 0.05$ adopted as the significance threshold. Feature variables exhibiting significant linear correlations with AGB were initially screened and retained as important inputs for subsequent model construction.

2.7. Random Forest Algorithm

The RF algorithm extends and generalizes traditional decision tree methods by integrating multiple decision trees into an ensemble framework [30]. It is well suited for the rapid and efficient processing of large-scale datasets and provides reliable measures of variable importance [32,33]. Building upon bagging-based ensembles that employ decision trees as base learners, RF additionally incorporates random feature selection during tree construction. Specifically, both training samples and feature subsets are randomly sampled from the original dataset to generate a collection of diverse decision trees. The predictions of these individual trees are subsequently aggregated, typically through averaging for regression tasks or majority voting for classification tasks, to produce the final prediction. As a result, RF regression models exhibit strong fitting capability while remaining relatively resistant to overfitting [32].

The key characteristics of the Random Forest model are summarized as follows. First, random sampling is employed during model training, whereby each decision tree is constructed using only a subset of the available samples and features, which enhances model diversity and reduces overfitting risk. Second, Random Forests have been widely applied to classification, regression, and feature selection tasks and have achieved robust performance across diverse application domains. Compared with many other machine learning algorithms, Random Forests exhibit strong robustness and generalization ability. Moreover, they are relatively easy to implement, apply, interpret, and tune, which has led to their extensive adoption in forest parameter estimation studies [34]. Their effectiveness in handling high-dimensional and imbalanced datasets, as well as their capability for variable selection, has been widely documented in previous research [32,35].

 doi:10.20944/preprints202603.2261.v1

2.8. Support Vector Machine Algorithm

SVM is a supervised learning algorithm grounded in statistical learning theory and the principle of structural risk minimization [36]. Its fundamental objective is to determine an optimal hyperplane that separates samples from different classes in the feature space for classification tasks while maximizing the minimum distance, or margin, between the hyperplane and the nearest data points. For nonlinear problems, SVM employs kernel functions to project the original data into a higher-dimensional feature space, where linear separability can be achieved, thereby effectively mitigating the so-called "curse of dimensionality."

A key advantage of SVM is its strong generalization capability, which makes it suitable for problems characterized by small sample sizes and high-dimensional feature spaces. This advantage arises from its strict adherence to the structural risk minimization principle, which aims to minimize training error (empirical risk) while simultaneously constraining model complexity. By balancing empirical risk and model capacity, SVM is designed to minimize an upper bound on the generalization error [36]. As a result, SVM effectively mitigates overfitting and has been shown to perform effectively in applications such as remote sensing inversion, where limited training samples are common. In this study, the regression variant of SVM, namely Support Vector Regression (SVR), is employed to model continuous variables such as AGB.

2.9. Pigeon-Inspired Optimization Algorithm

Pigeon-inspired optimization is a novel swarm intelligence–based optimization algorithm characterized by a simple operational mechanism and a limited number of control parameters. Compared with other swarm intelligence algorithms, PIO is relatively easy to implement and exhibits fast convergence behavior [37]. The algorithm was first proposed in 2014 by Duan Haibin et al., inspired by the collective homing behavior of pigeons [38]. Since its introduction, PIO and its variants have demonstrated strong performance in a range of complex optimization tasks, including parameter optimization, scheduling problems, and machine learning model tuning [15].

The PIO algorithm models pigeon homing behavior through two main components: the map-and-compass operator and the landmark operator. When pigeons are far from their destination, they rely on geomagnetic fields and the sun to determine flight direction; this behavior is simulated by the map-and-compass operator. As pigeons approach their destination, they increasingly depend on familiar landmarks for navigation, which is represented by the landmark operator. Accordingly, the map-and-compass operator is constructed based on geomagnetic and solar navigation principles, whereas the landmark operator is based on landmark-based guidance.

Pigeons are assumed to possess innate magnetic sensitivity, enabling them to perceive geomagnetic fields and adjust their flight direction according to the position of the sun during long-distance homing. As the flock approaches its destination, the influence of geomagnetic cues and solar navigation gradually diminishes. In the PIO framework, the solution space is defined as a D-dimensional space, where the position and velocity of the i-th pigeon ($i = 1, 2, \ldots, N$) represent candidate solutions. At each iteration, the positions and velocities of all pigeons are updated according to predefined update rules, as expressed in Equations (3) and (4) [38]:

$$V_{i(t)} = V_{i(t-1)} \cdot e^{\{-R \cdot t\}} + \{rand\} \cdot \left(X_g - X_{i(t-1)}\right) \quad (2)$$

$$X_{i(t)} = X_{i(t-1)} + V_{i(t)} \quad (3)$$

where R denotes the map-and-compass factor ($0 < R < 1$), X represents the global best position, t is the current iteration number, and *rand* is a uniformly distributed random number in the range [0, 1].

The landmark operator is established based on pigeons' navigation behavior when approaching their destination, where landmark guidance becomes more dominant than geomagnetic navigation. If pigeons are unfamiliar with the landmarks at their current location, they fly under the guidance of nearby pigeons. Once a landmark or familiar location is identified, pigeons rely on accumulated experience to guide their flight independently. In the landmark model, $N_p(t)$ denotes the number of

 doi:10.20944/preprints202603.2261.v1

pigeons in the t-th generation, and $X_{c(t)}$ represents the center position of the pigeon swarm. Assuming that each pigeon can fly directly toward the destination, the population size is updated as:

$$N_{\mathrm{p}}(t) = ceil[N_{\mathrm{p}}(t-1)/2] \quad (4)$$

The center position of the pigeon swarm is calculated as a fitness-weighted mean:

$$X_{c(t)} = [\sum(X_{\mathrm{i}}(t) \times \mathrm{fitness}(X_{\mathrm{i}}(t)))] \div [N_{\mathrm{p}} \times \sum \mathrm{fitness}(X_{\mathrm{i}}(t))] \quad (5)$$

Subsequently, the position of each pigeon is updated according to:

$$X_{i(t)} = X_{i(t-1)} + \{rand\} \cdot (X_{c(t)} - X_{i(t-1)}) \quad (6)$$

where fitness(X) represents the fitness value of each pigeon, reflecting the quality of the corresponding solution.

In this study, the PIO algorithm is employed to optimize the key hyperparameters of SVM, including the penalty parameter C and the kernel function parameter γ. By automatically searching for optimal parameter combinations, the PIO-SVM framework aims to enhance the accuracy and robustness of forest AGB estimation models [38].

Model accuracy assessment relies on appropriate evaluation metrics that effectively reflect both model fitting ability and predictive performance. In this study, three machine learning models were employed to estimate forest aboveground biomass in the Taishan forest region: PIO-SVM, SVM, and RF model. To quantitatively evaluate and compare the performance of these regression models, two widely used statistical metrics were used: the root mean square error (RMSE) and the coefficient of determination (R^2).

The available dataset was randomly divided into training and testing subsets at a ratio of 7:3. The evaluation metrics were calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

where n denotes the number of samples, y_i represents the observed AGB value, $\hat{y}_i$ is the model-predicted AGB value, and $\bar{y}$ is the mean of the observed AGB values. RMSE reflects the overall prediction error magnitude, while R^2 indicates the proportion of variance in AGB explained by the model.

3. Results

3.1. Correlation Analysis

After extracting a total of 105 remote-sensing derived feature variables, including 10 spectral bands, 6 vegetation indices, 80 texture features, 6 environmental factors, and 3 Tasseled Cap Transformations, Pearson correlation analysis was used to identify feature variables that exhibited significant correlations with forest AGB. A total of 17 feature variables were retained as inputs for subsequent machine learning modeling. The results indicate that the KNDVI shows the strongest correlation with AGB among all candidate variables, as summarized in Figure 4, Table 4.

 doi:10.20944/preprints202603.2261.v1

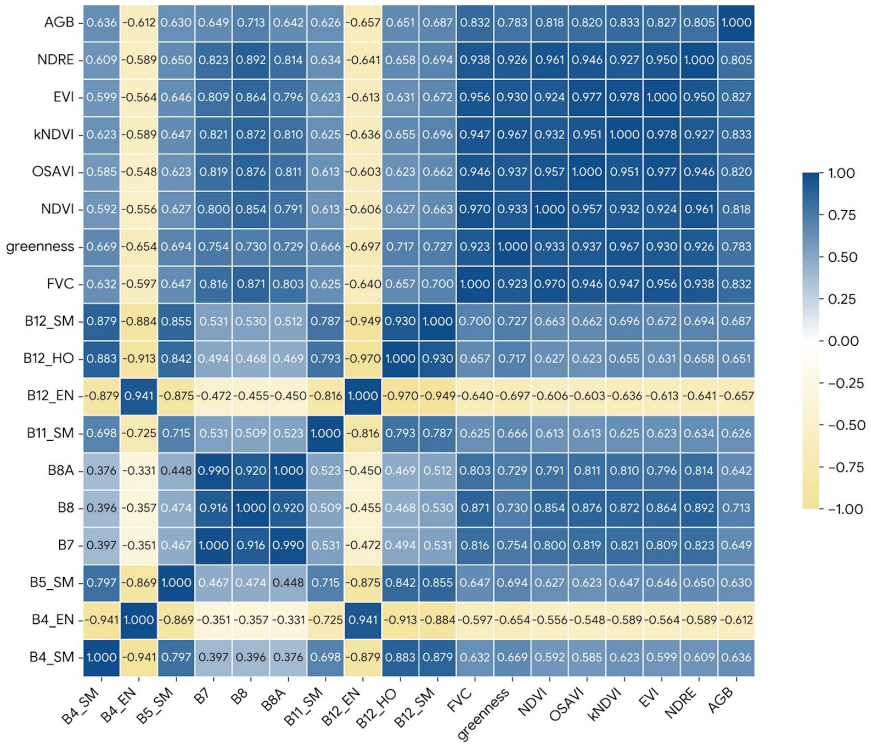


Figure 4. Correlation coefficient matrix between characteristic variables and AGB.

Table 4. Significant Correlation Coefficients for Remote Sensing Parameters.

Feature	B4_SM	B4_EN	B5_SM	B8	B8A
correlation	0.636***	-0.612***	0.63***	0.713***	0.642***
Feature	B11_SM	B12_EN	B12_SM	B12_HO	B7
correlation	0.626***	-0.657***	-0.687***	0.651***	0.649***
Feature	NDRE	OSAVI	EVI	FVC	greenness
correlation	0.805***	0.820***	0.827***	0.832***	0.783***
Feature	NDVI	KNDVI			
correlation	0.818***	0.833***			

Note: ***, **, and * denote significance levels of 1%, 5%, and 10%, respectively.

3.2. Evaluation of Model Accuracy

The selected model parameters and measured aboveground biomass data from 205 sample plots were imported into the MATLAB environment for model training and evaluation. The dataset was randomly divided into training and testing subsets at a ratio of 7:3. Scatter plots illustrating the relationships between measured and predicted AGB values for the different models are presented in Figure 5.

 doi:10.20944/preprints202603.2261.v1

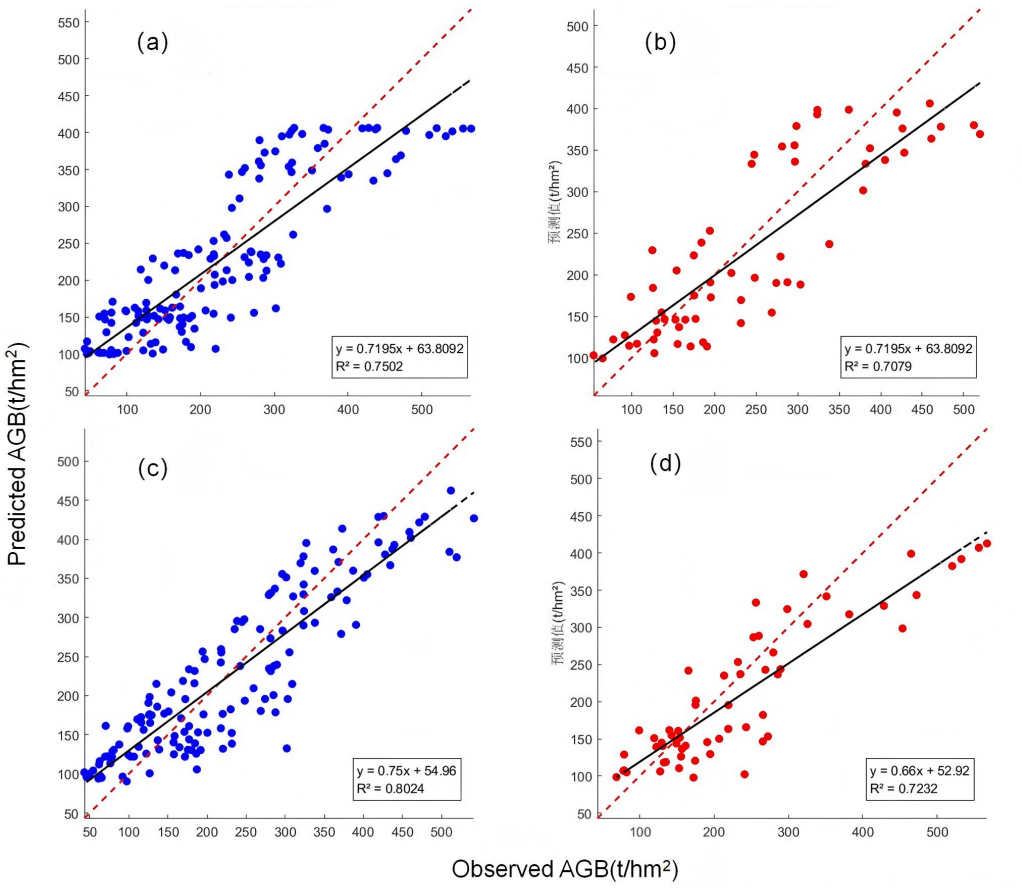


Figure 5. Scatter plot between predicted and observed AGB. (a) Modeling Set of RF; (b) Validation set of RF; (c) Modeling Set of SVM; (d) Validation set of SVM.

3.3. AGB Estimation Based on PIO-SVM

Forest aboveground biomass estimation was conducted using SVM optimized by the PIO algorithm. The SVM model was trained using the optimized penalty parameter (C) and insensitive loss parameter (ε), with the training and testing datasets randomly divided at a ratio of 7:3. The estimation results obtained using the PIO-SVM model are presented in Figure 6.

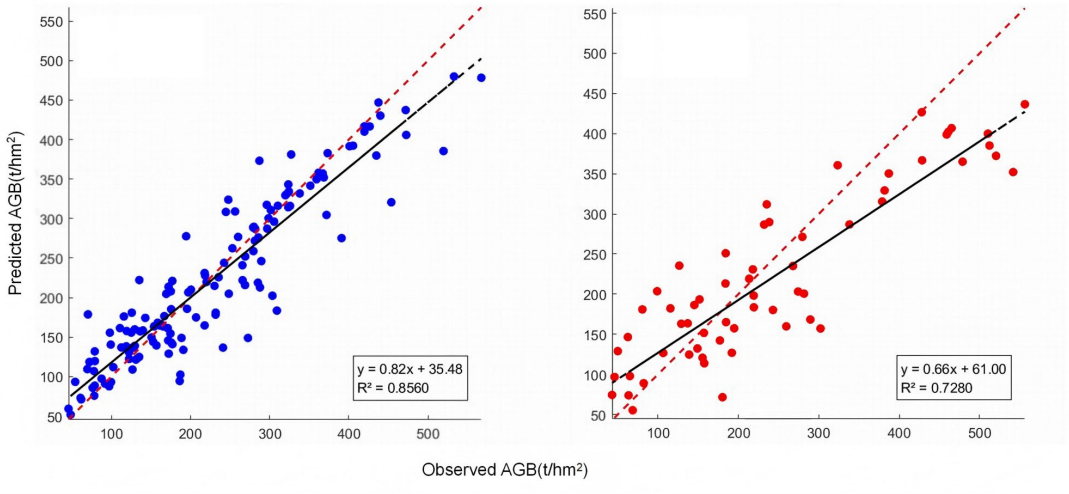


Figure 6. Scatter plots of the predicted AGB by PIO-SVM model.

 doi:10.20944/preprints202603.2261.v1

3.4. Model Accuracy Evaluation

In this study, three machine learning algorithms—RF, SVM, and PIO-SVM—were employed to construct forest aboveground biomass estimation models for the Mount Tai region. All models were developed using Sentinel-2 multispectral imagery in combination with an optimized set of environmental variables. The quantitative accuracy assessment results for the three models are summarized in Table 5.

Table 5. Accuracy Evaluation Results for Different Models.

Model	Modeling Set		Validation Set	
	R^2	RMSE(t/ha)	R^2	RMSE(t/ha)
RF	0.75	62.41	0.70	65.19
SVM	0.80	54.85	0.72	66.25
PIO-SVM	0.85	46.12	0.73	62.19

As shown in Table 5, the PIO-SVM model demonstrates the best overall fitting performance and predictive accuracy among the evaluated models. On the training dataset, the PIO-SVM model achieves the highest coefficient of determination (R^2 = 0.85) and the lowest root mean square error (RMSE = 46.12). On the testing dataset, it also outperforms the other models, with an R^2 of 0.73 and an RMSE of 62.19. The standard SVM model exhibits moderate performance, with a training R^2 of 0.80 and RMSE of 54.85, and a testing R^2 of 0.72 with an RMSE of 66.25. The Random Forest model shows comparatively lower accuracy, yielding a training R^2 of 0.75 and RMSE of 62.41, and a testing R^2 of 0.70 with an RMSE of 65.19. Overall, all three models achieve coefficients of determination exceeding 0.70 on both the training and testing datasets, indicating satisfactory estimation capability for forest AGB. Among them, the PIO-SVM model exhibits superior generalization performance and reduced prediction error, highlighting the effectiveness of PIO in enhancing SVM-based biomass estimation under complex mountainous conditions.

3.5. Forest Aboveground Biomass Estimation Mapping

The optimal model obtained using the PIO-SVM was applied to estimate forest aboveground biomass across the entire study area. The results indicate that forest AGB is predominantly concentrated in the central and northern regions of the study area, where forest cover is relatively dense. Biomass density exhibits a clear spatial gradient, gradually increasing from the peripheral areas toward the interior of the region. This spatial pattern is highly consistent with field survey observations, demonstrating the reliability of the PIO-SVM model for large-scale forest AGB mapping. The spatial distribution of estimated AGB is illustrated in Figure 7.

 doi:10.20944/preprints202603.2261.v1

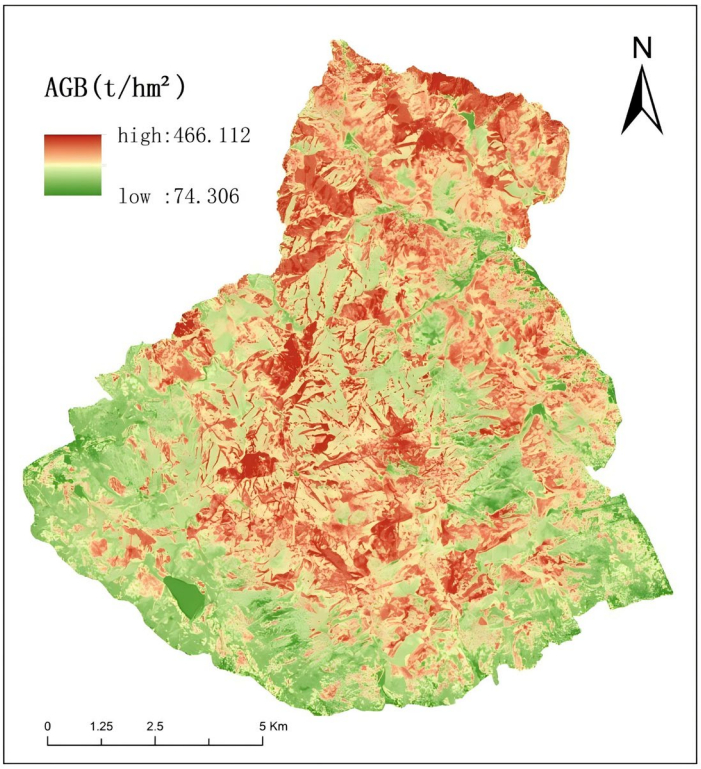


Figure 7. Spatial distribution of forest AGB estimated using the PIO-SVM model in the study area.

4. Discussion

4.1. Performance Comparison of Different Machine Learning Models

Different machine learning models exhibit varying capacities for handling complex data structures and environmental heterogeneity. In the mountainous forest environment of Mount Tai, the three models examined in this study—RF, SVM, and PIO-SVM—showed noticeable differences in their performance for AGB estimation. Among them, the PIO-SVM model achieved higher accuracy on both the training and testing datasets, with R^2 values of 0.85 and 0.73, respectively, together with lower RMSE values than those obtained using the RF and conventional SVM models.

The superior performance of the PIO-SVM model is primarily attributed to the global optimization capability of the PIO algorithm in tuning key SVM hyperparameters, including the penalty coefficient C and the kernel parameter γ. Conventional SVM models are highly sensitive to hyperparameter selection, and suboptimal configurations may result in convergence to local optima or reduced generalization performance. By contrast, RF models reduce variance through ensemble learning and exhibit strong robustness to noise with minimal parameter tuning; however, their ability to capture highly complex nonlinear relationships may be constrained in heterogeneous forest environments. The PIO algorithm employed in this study balances global exploration and local exploitation of the parameter space by simulating the two-stage homing behavior of pigeons, represented by a global map-and-compass operator and a local landmark operator. This optimization strategy enables the SVM to identify parameter combinations closer to the global optimum, thereby improving predictive accuracy, convergence stability, and robustness in modeling complex nonlinear relationships between multi-source remote sensing features and forest AGB in mountainous terrain. Consequently, the PIO-SVM model achieves a favorable balance between accuracy and stability under complex terrain conditions, demonstrating the effectiveness of PIO for enhancing traditional machine learning models.

 doi:10.20944/preprints202603.2261.v1

4.2. Consistency Analysis and Comparison with Previous Studies

The findings of this study are generally consistent with previous research emphasizing the role of optimization strategies in improving forest AGB estimation accuracy. Existing studies have demonstrated that model performance can be enhanced either by refining feature selection procedures or by improving model parameterization, particularly when high-dimensional remote sensing datasets are involved.

For example, Wang et al. constructed an SVR model based on Lasso-based feature selection, with an emphasis on reducing feature redundancy and improving generalization performance [39]. Although both studies aim to optimize model performance, this study instead emphasizes parameter space optimization through adaptive hyperparameter tuning using the PIO algorithm. While both feature selection and parameter optimization can substantially enhance SVR performance, the PIO-SVM model demonstrates superior performance in the highly heterogeneous and nonlinear mountainous environment of Mount Tai. This finding suggests that in regions characterized by strong feature heterogeneity and complex nonlinear relationships, hyperparameter optimization may be as important as, or even more critical than, feature selection alone.

In contrast, Luo et al. employed regression kriging to correct spatially autocorrelated residuals from Random Forest predictions, thereby improving spatial mapping accuracy, particularly by mitigating overestimation and underestimation at extreme values [40]. In the present study, residual-based spatial correction was not conducted. Instead, terrain-related factors were considered at an earlier stage of analysis. Variables such as slope and terrain wetness index were included as candidate factors during feature screening to examine their potential associations with forest growth conditions. The inclusion of these variables contributes to more realistic spatial patterns in the AGB estimates produced by the PIO-SVM model, particularly in areas with pronounced terrain variation. In general, post-processing spatial correction methods and feature before or during the process and parameter optimization approaches should be considered as complementary rather than mutually exclusive. Integrating the parameter optimization strengths of the PIO-SVM framework with spatial residual correction methods such as regression kriging represents a promising avenue for further enhancing forest AGB estimation accuracy.

4.3. Limitations and Outlook

Although the PIO-SVM model shows relatively good performance in this study, several limitations should be acknowledged, mainly associated with data sources, terrain conditions, and sampling constraints. First, optical remote sensing data are subject to spectral saturation in areas with high biomass. Under dense canopy conditions, the reflectance signals recorded by sensors such as Sentinel-2 tend to saturate in the visible and near-infrared bands, reducing sensitivity to further increases in AGB. As a result, biomass in high-density forest stands may be underestimated, a phenomenon that has also been reported in previous studies. Second, the complex terrain and heterogeneous forest composition of the Mount Tai region introduce additional uncertainty into the modeling process. Large elevation differences lead to spatial variability in illumination and moisture conditions, which in turn affect vegetation growth and spectral responses. In addition, coniferous, broadleaf, and mixed forests exhibit different canopy structures and seasonal characteristics, making it difficult to establish a single relationship that performs equally well across all forest types. Although texture features and topographic variables were examined during feature screening, estimation errors still occur in areas with steep slopes or structurally complex mixed forests. Third, the spatial distribution of field samples imposes limitations on model generalization. Due to accessibility constraints, sample plots are unevenly distributed, with fewer observations located in high-elevation or steep terrain. Furthermore, some forest age classes are underrepresented in the dataset, which may affect the reliability of AGB estimates at both the stand and regional scales. Overall, these limitations indicate that further improvements could be achieved by integrating complementary data sources, such as active remote sensing data, expanding field sampling coverage, and refining modeling strategies for heterogeneous mountainous forests.

 doi:10.20944/preprints202603.2261.v1

To address these limitations, future research should pursue several complementary directions. At the data level, multisource data fusion integrating optical, radar, and LiDAR data offers a promising approach for mitigating spectral saturation effects. Radar data such as Sentinel-1 SAR or ALOS-2 PALSAR-2 are sensitive to vegetation vertical structure, while airborne or satellite LiDAR can directly capture three-dimensional forest structural information, thereby enabling the construction of feature sets with stronger physical interpretability and improved discrimination in high-biomass regions. At the methodological level, further integration of advanced modeling approaches is warranted. Combining PIO algorithms with deep learning architectures (e.g., convolutional neural networks or Transformer-based models) may facilitate the automated extraction of deeper spatial–spectral features. In addition, incorporating spatial statistical techniques—such as coupling PIO-SVM with regression kriging for residual correction—could establish a hybrid framework integrating intelligent parameter optimization with spatial error correction, thereby reducing prediction bias and improving spatial continuity.

Finally, to enhance the ecological interpretability and applicability of AGB estimation models, future studies should further integrate ecological processes by incorporating multidimensional factors such as forest age, species composition, soil properties, climate variability, and human disturbances. This transition from purely data-driven modeling toward a data–mechanism synergistic framework would not only improve estimation accuracy but also support long-term monitoring of forest carbon sinks, ecosystem dynamics, and the scientific assessment of regional carbon neutrality strategies.

5. Conclusions

A robust framework for estimating aboveground biomass (AGB) in the Mount Tai region, a temperate mountainous forest area in eastern China, was established by integrating multisource remote sensing data and intelligent parameter optimization. The performance of the proposed PIO-SVM model was evaluated against standard SVM and Random Forest (RF) models. The main conclusions can be summarized as follows.

(1) Feature screening based on Pearson correlation analysis identified spectral, red-edge, vegetation index, texture, and topographic variables as key contributors to the spatial variability of forest AGB in mountainous environments. Among these, FVC, KNDVI, and NDRE exhibited the strongest correlations with field-measured AGB, highlighting their utility for biomass estimation in complex terrain.

(2) The overall performance of the PIO-SVM model was evaluated relative to standard SVM and Random Forest (RF) models, demonstrating superior predictive accuracy and robustness. Adaptive optimization of SVM hyperparameters enabled the PIO-SVM model to achieve higher fitting accuracy and more stable performance across both training and testing datasets, emphasizing the critical role of parameter optimization under heterogeneous topographic conditions.

(3) The spatial distribution of AGB derived from the optimal PIO-SVM model showed pronounced heterogeneity across the study area, with higher biomass values concentrated in the central and northern forested regions of Mount Tai. These spatial patterns were in close agreement with field survey observations, confirming the reliability of the proposed approach for generating accurate AGB maps in mountainous forest ecosystems. The results demonstrate that the PIO algorithm enhances the performance of support vector machine–based models by enabling more effective parameter selection in complex modeling scenarios. The proposed framework provides a practical and reliable approach for estimating forest AGB in heterogeneous mountainous regions and facilitates regional-scale biomass monitoring.

Author Contributions: Conceptualization, X.K. and L.W.; methodology, X.K.; software, X.K. and X.L.; validation, C.Q. and D.C.; investigation, X.K., C.Q. and X.M.; writing—original draft preparation, X.K.; writing—review and editing, X.K., L.W., C.C. and X.Z.; funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202603.2261.v1

Funding: This research was funded by Natural Science Foundation of Shandong Province, grant number ZR2023MD083; National Natural Science Foundation of China, grant number 42171378.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to thank the kind help of the editor and the reviewers for improving the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xu, W.Y.; Jin, X.B.; Yang, X.H.; et al. Spatial grid-based estimation of forest vegetation biomass in China. *J. Nat. Resour*. **2018**, 33, 1725–1741. https://doi.org/10.31497/zrzyxb.20170802
2. Kumar, L.; Mutanga, O. Remote sensing of above-ground biomass. *Remote Sens*. **2017**, 9, 935. https://doi.org/10.3390/rs9090935
3. Saatchi, S.S.; Harris, N.L.; Brown, S.; et al. Benchmark map of forest carbon stocks in tropical regions across three continents. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, 108, 9899–9904. https://doi.org/10.1073/pnas.1019576108
4. Luo, Y.J.; Zhang, X.Q.; Wang, X.K.; et al. Methods for estimating forest biomass and research progress. *For. Sci.* **2009**, 45, 129–134. https://doi.org/10.11707/j.1001-7488.20090823
5. Breiman, L. Random forests. *Mach. Learn.* **2001**, 45, 5–32. https://doi.org/10.1023/A:1010933404324
6. Zou, W.T.; Zeng, W.S. Advances in forest biomass and carbon stock estimation. *World For. Res.* **2025**, 38, 32–38. https://doi.org/10.13348/j.cnki.sjlyyj.2025.0031.
7. Tian, L.; Wu, X.; Tao, Y.; et al. Review of remote sensing-based methods for forest aboveground biomass estimation: Progress, challenges, and prospects. *Forests* **2023**, 14, 1086. https://doi.org/10.3390/f14061086
8. Hengl, T.; Heuvelink, G.B.M.; Rossiter, D.G. About regression-kriging: From equations to case studies. *Comput. Geosci.* **2007**, 33, 1301–1315. https://doi.org/10.1016/j.cageo.2007.05.001
9. Foody, G.M.; Boyd, D.S.; Cutler, M.E.J. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. *Remote Sens. Environ.* **2003**, 85, 463–474. https://doi.org/10.1016/S0034-4257(03)00039-7
10. Zhang, B.H.; Zhang, L.; Xie, D.; et al. Application of synthetic NDVI time series blended from Landsat and MODIS data for grassland biomass estimation. *Remote Sens.* **2015**, 8, 10. https://doi.org/10.3390/rs8010010
11. Dang, A.T.N.; Nandy, S.; Srinet, R.; et al. Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. *Ecol. Inform.* **2019**, 50, 24–32. https://doi.org/10.1016/j.ecoinf.2018.12.010
12. Wang, Y.F.; Sun, Y.J.; Guo, X.Y. Standing biomass model of Pinus massoniana based on BP neural network. *J. Beijing For. Univ.* **2013**, 35, 17–21. https://doi.org/10.13332/j.1000-1522.2013.02.018
13. Lin, Z.; Wu, C.Z.; Hong, W.; et al. Harvesting model of artificial cypress forests based on BP neural network and support vector machine. *J. Beijing For. Univ.* **2015**, 37, 42–47. https://doi.org/10.13332/j.cnki.jbfu.2015.01.008
14. Zhang, P.C.; Liang, Y.; Liu, B.; et al. Remote sensing estimation of forest above-ground biomass on the Qinghai–Tibet Plateau based on random forest models. *J. Ecol.* **2023**, 42, 415–424. https://doi.org/10.13292/j.1000-4890.202302.007
15. Liu, H.; Yan, X.; Wu, Q. An improved pigeon-inspired optimisation algorithm and its application in parameter inversion. *Symmetry* **2019**, 11, 1291. https://doi.org/10.3390/sym11101291
16. Liu, S.T.; Gao, P.; Liu, P.W.; et al. Assessment of ecosystem services and their value in the Taishan forest ecosystem. *Acta Ecol. Sin.* **2017**, 37, 3302–3310. https://doi.org/10.5846/stxb201511302396
17. Guidelines on carbon accounting and monitoring for afforestation project. Available online: https://std.samr.gov.cn/hb/search/stdHBDetailed?id=8B1827F1BB6BBB19E05397BE0A0AB44A.
18. Yu, S.; Ye, Q.; Zhao, Q.; et al. Effects of driving factors on forest aboveground biomass in China's Loess Plateau using spatial regression models. *Remote Sens.* **2022**, 14, 2842. https://doi.org/10.3390/rs14122842
19. Hao, Q.; Huang, C. Review of forest aboveground biomass estimation based on remote sensing data. *Chin. J. Plant Ecol.* **2023**, 47, 1356–1374. https://doi.org/10.17521/cjpe.2023.0008

20. Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, SMC-3, 610–621. https://doi.org/10.1109/TSMC.1973.4309314
21. Shen, Q.; Guo, L.; Zhang, B. Integration of spectral and textural information for improving land use/land cover classification. *Int. J. Remote Sens.* **2020**, 41, 3281–3299. https://doi.org/10.1080/01431161.2019.1701210
22. Jiang, F.; Kutia, M.; Ma, K.; et al. Estimating aboveground biomass of coniferous forest using spectral variables, land surface temperature and soil moisture. *Sci. Total Environ.* **2021**, 785, 147335. https://doi.org/10.1016/j.scitotenv.2021.147335
23. Yuan, F.; Bauer, M.E. Comparison of impervious surface area and NDVI as indicators of surface urban heat island effects. *Remote Sens. Environ.* **2007**, 106, 375–386. https://doi.org/10.1016/j.rse.2006.09.003
24. Dong, T.; Liu, J.; Shang, J.; et al. Assessment of red-edge vegetation indices for crop leaf area index estimation. *Remote Sens. Environ.* **2019**, 222, 133–143. https://doi.org/10.1016/j.rse.2018.12.032
25. Purevdorj, T.S.; Tateishi, R.; Ishiyama, T.; Honda, Y. Relationships between percent vegetation cover and vegetation indices. *Int. J. Remote Sens.* **1998**, 19, 3519–3535. https://doi.org/10.1080/014311698213795
26. Sharifi, A.; Felegari, S. Remotely sensed normalized difference red-edge index for rangeland biomass estimation. *Aircr. Eng. Aerosp. Technol.* **2023**, 95, 1128–1136. https://doi.org/10.1108/AEAT-07-2022-0199
27. Rondeaux, G.; Steven, M.; Baret, F. Optimization of soil-adjusted vegetation indices. *Remote Sens. Environ.* **1996**, 55, 95–107. https://doi.org/10.1016/0034-4257(95)00186-7
28. Guo, B.; Xu, M.; Zhang, R.; Lu, M. Dynamic monitoring of rocky desertification using Sentinel-2 images and KNDVI. *Geomatics Nat. Hazards Risk* **2024**, 15, 1–18. https://doi.org/10.1080/19475705.2024.2399659
29. Manjunath, B.S.; Ma, W.Y. Texture features for browsing and retrieving large image data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, 33, 117–128. https://doi.org/10.1109/34.531803
30. Shi, T.; Xu, H. Derivation of tasseled cap transformation coefficients for Sentinel-2 MSI data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, 12, 4038–4048. https://doi.org/10.1109/JSTARS.2019.2938388
31. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random forests. In *Ensemble Machine Learning: Methods and Applications*; Springer: New York, NY, USA, **2012**; pp. 157–175. https://doi.org/10.1007/978-1-4419-9326-7_5
32. Belgiu, M.; Drăguț, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011
33. Houghton, R.A.; Butman, D.; Bunn, A.G.; et al. Mapping Russian forest biomass with satellite and inventory data. *Environ. Res. Lett.* **2007**, 2, 045032. https://doi.org/10.1088/1748-9326/2/4/045032
34. Xiong, K.; Li, S.F.; Xing, Y.J. Remote sensing estimation of aboveground biomass in forest lands of Yuanling County, Hunan Province. *Cent. South For. Surv. Plan.* **2024**, 43, 44–50. https://doi.org/10.16166/j.cnki.cn43-1095.2024.01.010
35. Scornet, E.; Biau, G.; Vert, J.-P. Consistency of random forests. *Ann. Stat.* **2015**, 43, 1716–1741. https://doi.org/10.1214/15-AOS1321
36. Ding, S.F.; Qi, B.J.; Tan, H.Y. Review of support vector machine theory and algorithms. *J. Univ. Electron. Sci.* Technol. China **2011**, 40, 2–10.
37. Wei, J.Y. Research on support vector machine parameter optimization based on pigeon flock algorithm. Master's Thesis, Hebei University of Geosciences, Shijiazhuang, China, **2019**.
38. Duan, H.; Qiao, P. Pigeon-inspired optimization: A new swarm intelligence optimizer for air robot path planning. *Int. J. Intell. Comput. Cybern.* **2014**, 7, 24–37. https://doi.org/10.1108/IJICC-02-2014-0005
39. Wang, P.; Tan, S.; Zhang, G.; et al. Remote sensing estimation of forest aboveground biomass based on Lasso-SVR. *Forests* **2022**, 13, 1597. https://doi.org/10.3390/f13101597
40. Luo, Y.N.; et al. Estimation of forest aboveground biomass in complex mountainous areas using regression kriging. *Forests* **2024**, 15, 1734. https://doi.org/10.3390/f15101734