sustainability

MDPI

Article

Biomass Estimation of Subtropical Arboreal Forest at Single Tree Scale Based on Feature Fusion of Airborne LiDAR Data and Aerial Images

Min Yan [1], Yonghua Xia [1,2,*], Xiangying Yang [3], Xuequn Wu [1], Minglong Yang [1,2], Chong Wang [4], Yunhua Hou [2] and Dandan Wang [2]

[1] Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650093, China
[2] City College, Kunming University of Science and Technology, Kunming 650051, China
[3] Faculty of Public Administration, Yunnan University of Finance and Economics, Kunming 650221, China
[4] Kunming Survey, Design and Research Institute Co., Ltd. of China Power Construction Group, Kunming 650200, China
* Correspondence: 20040063@kust.edu.cn

Abstract: Low-cost UAV aerial photogrammetry and airborne lidar scanning have been widely used in forest biomass survey and mapping. However, the feature dimension after multisource remote sensing fusion is too high and screening key features to achieve feature dimension reduction is of great significance for improving the accuracy and efficiency of biomass estimation. In this study, UAV image and point cloud data were combined to estimate and map the biomass of subtropical forests. Firstly, a total of 173 dimensions of visible light vegetation index, texture, point cloud height, intensity, density, canopy, and topographic features were extracted as variables. Secondly, the Kendall Rank correlation coefficient and permutation importance (PI) index were used to identify the key features of biomass estimation among different tree species. The random forest (RF) model and XGBoost model finally were used to compare the accuracy of biomass estimation with different variable sets. The experimental results showed that the point cloud height, canopy features, and topographic factors were identified as the key parameters of the biomass estimate, which had a significant influence on the biomass estimation of the three dominant tree species in the study area. In addition, the differences in the importance of characteristics among the tree species were discussed. The fusion features combined with the PI index screening and RF model achieved the best estimation accuracy, the R2 of 0.7356, 0.8578, and 0.6823 were obtained for the three tree species, respectively.

Keywords: multi-source remote sensing fusion; feature screening; single tree scale; subtropical arboreal forest; estimation of biomass

check for updates

Citation: Yan, M.; Xia, Y.; Yang, X.; Wu, X.; Yang, M.; Wang, C.; Hou, Y.; Wang, D. Biomass Estimation of Subtropical Arboreal Forest at Single Tree Scale Based on Feature Fusion of Airborne LiDAR Data and Aerial Images. *Sustainability* **2023**, *15*, 1676. https://doi.org/10.3390/su15021676

Academic Editor: Eben Broadbent

Received: 27 November 2022
Revised: 8 January 2023
Accepted: 12 January 2023
Published: 15 January 2023

1. Introduction

The estimation of forest biomass, as an important index to measure the growth potential and carbon sequestration capacity of forests, and its mapping are of great significance in forest resource surveys nowadays. How to estimate and map it quickly and accurately is one of the urgent problems to be solved [1–3]. Traditional biomass survey methods, based on field measurements, usually require a lot of time and high labor costs and have difficulties in achieving highly precise and large-scale biomass mapping [4–8]. With the development of remote sensing, UAV remote sensing enables the rapid, non-destructive, and accurate mapping of biomass with the advantages of flexible operation, high resolution, and non-contact sensing [9,10].

Vegetation spectral information, such as the vegetation index and image texture, can fully express leaf color, vegetation richness, and health status [11,12]. In recent years, some studies have applied the visible light spectrum [13], image texture [14,15], and visible vegetation index [16–18] to estimate vegetation biomass by using the practical and inexpensive measurement of UAVs equipped with cameras. Both Wang Li et al. (2016) [19] and Bo Li et al. (2020) [20] have used vegetation index (VI) features and vegetation height-related variables obtained using aerial photography to estimate aboveground biomass and forecast the yield of maize and potato, respectively. Yinuo Liu et al. (2019) [21] have used vegetation index and texture (Gray Level Co-occurrence Matrix, GLCM) of multi-spectral images for biomass estimation of rapeseed in winter.

Airborne point cloud data can easily extract features such as the vegetation height, density and topography, vertical structure, and canopy area [22–24] easily, which complement the vegetation index and texture features of images. It has been widely introduced into biomass estimation studies. Shengli Tao et al. (2014) [25] have obtained the single-tree canopy volume based on point cloud clustering segmentation and used it for the accurate estimation of aboveground biomass (AGB) in forests. Sami Ullah et al. (2017) [26] have extracted forest height- and density-related variables from airborne laser point clouds and aerial photographic point clouds to estimate forest timber stock. Linghan Gao et al. (2022) [27] have estimated the aboveground biomass of plantations of different tree species by extracting variables related to tree canopies, topography, point cloud height, and density.

In the study of biomass estimation based on multi-source remote sensing, the traditional linear parameter model has problems of nonlinearity and multicollinearity due to its limited statistical assumptions. As a result, non-parametric modeling methods are widely used in biomass estimation research, which solves the problems of high dimensionality, high redundancy, and small sample sizes in multi-source remote sensing data [28–30]. Coeli M. Hoover et al. (2018) [31] compared random forest (RF) with the traditional biomass estimation methods and proved that the estimate from RF was better than the general estimate by using the average canopy height and cross-sectional area, as proposed by G.P. Asner et al. (2011) [32]. In addition, Yue Zhang et al. (2021) [33] used the hyper-spectral narrowband vegetation index and crop height from UAVs to estimate maize bio-mass and obtained better results in estimating maize biomass using the XGBoost model than with the stepwise regression and RF regression models.

In this study, the visible vegetation index and texture features, airborne lidar point cloud intensity, density, height, and other features, as well as the canopy and topographic factors, were fused as input variables of the model. By comparing different variable combinations and feature selection methods, the biomasses of the three dominant tree species in the study area were estimated by using RF and XGBoost classical models. Additionally, the differences in the characteristics and importance of the various tree species were analyzed. The main objectives of this study are: (1) to determine the key characteristic parameters of subtropical tree biomass regression based on UAV multi-source remote sensing fusion (image and point cloud); (2) to compare the key characteristics of biomass estimation among different tree species; and (3) to realize the accurate estimation and mapping of single-tree scale biomass of arboreal forest by filtrating the non-important features and provide a reference for the research on multi-source remote sensing fusion and accurate biomass estimation.

2. Materials and Methods

2.1. Study Area

The study area is located in the Kuandiba forest area (24°43′–24°56′ N, 102°28′–102°38′ E) of the Haikou Forest Farm, Xishan District, Kunming City, Yunnan Province. It demonstrates the mountain topography of the Central Yunnan Plateau, with a "lake plateau" landform and subtropical monsoon climate. The average altitude is 1900–2200 m. The average annual temperature is 14.6 °C, with the highest of 34.4 °C and the lowest of −7.8 °C. The average annual rainfall is 909.7 mm, which happens in June and July mostly. The forest coverage rate in the study area was 80.46% and the main tree types were: Pinus Armandii.Franch, *Pinus Yunnanensis, Sabina Chinensis, Cupressus Lusitanica, Alnus Japonica*

Steud, Eucalyptus robusta Smith, Eucommia Ulmoides Oliver, and *China fir*. The Figure 1 has showed the location of the study area, the point cloud data and orthophoto image.

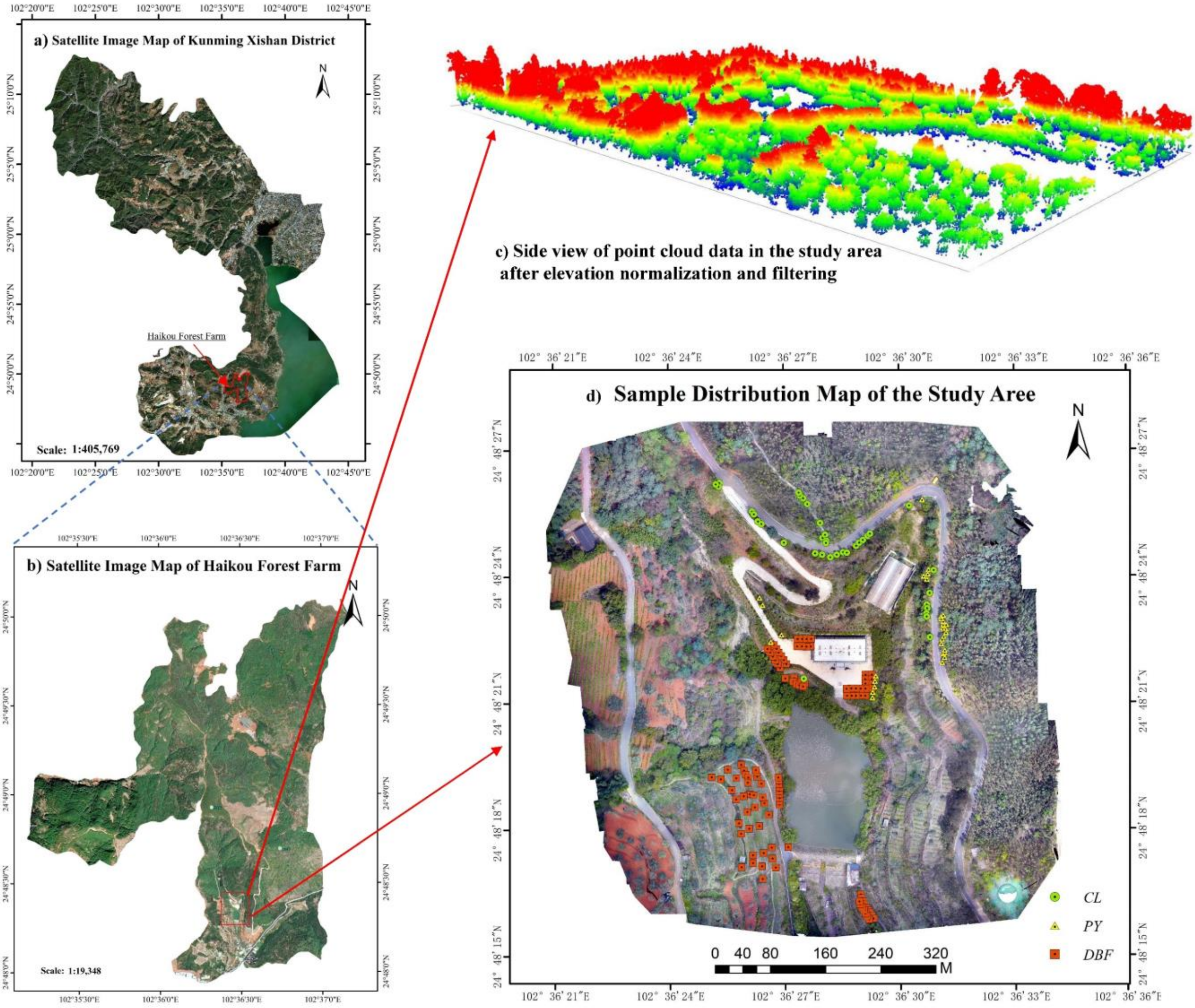

Figure 1. Location of the study area (**a**) is the location of Xishan District. (**b**) is the Haikou Forest Farm. (**c**) is the side view of point cloud data in the study area after elevation normalization and filtering. (**d**) is the UAV orthophoto image and sample distribution in study area.

2.2. Field Data Collection

The tree diameter at breast height (DBH) was obtained by measuring the circumference of the tree at 1.3 m manually. The heights of the trees were measured individually with their location recorded. The experiment was carried out on the three types of trees, namely, *Cupressus lusitanica (CL)*, *Pinus yunnanensis (PY)*, and *Deciduous Broad-leaved Forest (DBF)*. These abbreviations will be used extensively to represent the names of the three tree species in the below text. A total of 174 samples of the three tree types were measured. The Deciduous Broad-Leaved Forest samples mainly included *Olea europaea, Eucommia Ulmoides Oliver, Cinnamomum camphora (L.) Presl.*, *Metasequoia glyptostroboides*, and *Eucalyptus species*, which were uniformly classified as the *DBF* due to the lack of its relevant regional biomass equation or the small quantity. The sample sizes of the trees species are provided in Table 1 and the tree height and DBH distribution of each sample are shown in Figure 2.

Table 1. Sample size.

Species	*CL*	*PY*	*DBF*	Total
Quantity	36	31	107	174

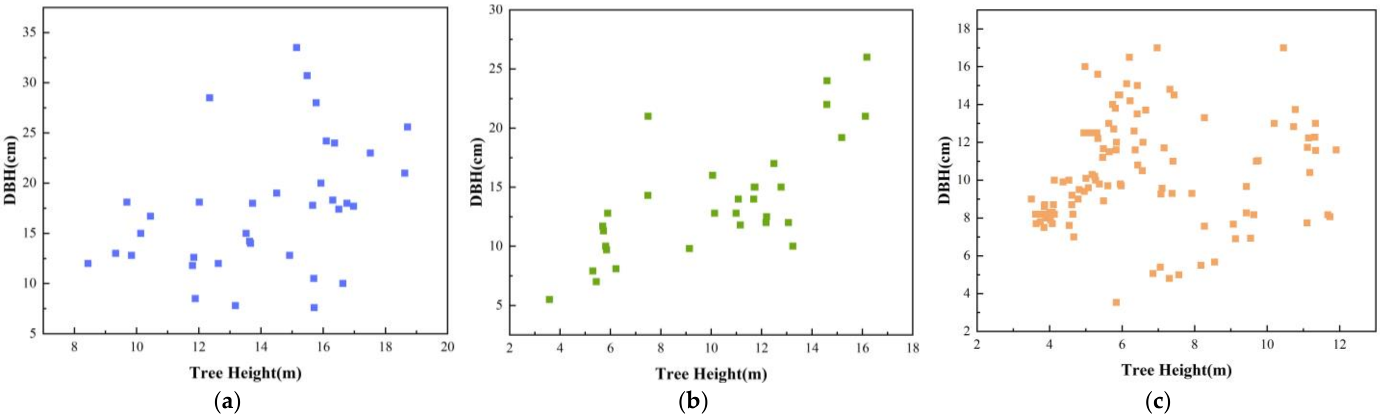

Figure 2. The tree height and DBH distribution map of samples. (**a**) *Cupressus lusitanica*, (**b**) *Pinus yunnanensis*, and (**c**) *Deciduous Broad-leaved Forest*.

The true biomass values of our experimental samples were obtained by using the binary biomass allometric equation of the tree height and DBH in the studies of Pan et al. [34] and Zhou et al. [35]. The application of these equations is limited to China (subtropical *DBF*) or the southwestern provinces of China (*CL*, *PY*). The allometric equation used in the study is shown in Table 2.

Table 2. Biomass allometric equation.

Species	Class	Equation	R^2	Reference
Cupressus lusitanica	Trunk	$W = 38.57447 \times (D2H)^{0.88545}$	0.99	[34]
	Branch	$W = 1.42899 \times (D2H)^{1.20845}$	0.98	
	Leaf	$W = 6.00855 \times (D2H)^{0.99494}$	0.99	
	Root	$W = 4.23742 \times (D2H)^{0.93532}$	0.99	
Pinus yunnanensis	Trunk	$W = 0.009 \times (D2H)^{1.044}$	0.98	[35]
	Branch	$W = 0.0008 \times (D2H)^{1.151}$	0.92	
	Leaf	$W = 0.006 \times (D2H)^{0.853}$	0.98	
	Root	$W = 0.009 \times (D2H)^{0.971}$	0.99	
Deciduous Broad-leaved Forest	Trunk	$W = 0.0263 \times (D2H)^{0.9695}$	0.98	[35]
	Branch	$W = 0.0232 \times (D2H)^{0.8055}$	0.97	
	Leaf	$W = 0.0075 \times (D2H)^{0.8015}$	0.96	
	Root	$W = 0.0381 \times (D2H)^{0.762}$	0.94	

Note: Among them, the tree height and DBH of *Cupressus lusitanica* are in meters. The other two types of tree height are in meters and the DBHs are in centimeters. In addition, the biomass is measured in kilograms.

2.3. UAV Data Collection

A DJI M300 (SZ DJI Technology Co., Ltd.; Shenzhen, China) multi-rotor unmanned aerial vehicle (UAV), equipped with an AA450 Lidar sensor (Shanghai Huace Navigation Technology Ltd.; Shanghai, China) and a Zenmuse P1 camera (SZ DJI Technology Co., Ltd.; Shenzhen, China), was used in this experiment. The technical parameters of the equipment are shown in Table 3. The flight height of the UAV is 75 m, the route overlap rate of aerial photography is 80%, and the side overlap rate is 70%.

Table 3. Equipment technical parameters.

Name	Parameters
AA450 airborne lidar sensor	Scanning frequency: 240,000 points/s (single echo), 720,000 points/s (triple echo); maximum range: 450 m; field angle: 70.4°(Horizontal) × 4.5°(Vertical)
Zenmuse P1 camera	Focal length: 24 mm; CCD size: 35.9 × 24 mm; Effective pixels: 45 million

2.4. Data Preprocessing

2.4.1. Preprocessing of Image and Point Cloud

The Pix-4D 4.5.6 software was used to process the original images collected by the UAVs with orthophotos generated in the *.tif format. The spatial resolution of the orthophoto results was 5 cm. The georeference of these images was carried out through 5 ground control points.

The airborne Lidar point cloud was spliced using CHC Navigation Co-Pre software. The UAV RTK differential post-processing was carried out using the base station at the ground control point (GCP). The result density was 382 points per square meter. The coordinates of the GCP and image phase control point were measured using the GPS equipment of the Galaxy-1 (South Surveying and Mapping Technology Co., LTD., Guangzhou, China). The point cloud data denoising, filtering, CHM production, and other processes of preprocessing were implemented using Lidar360 software (Beijing Green Valley Technology Co., LTD., Beijing, China). Lidar 360 is an efficient point cloud post-processing software, including a variety of point cloud processing tools, that is often used for the visual editing of point cloud data and the production of various geospatial products. In addition, it has been widely used in forestry point cloud data processing.

2.4.2. Segmentation in Single Tree Scale and Classification of Dominant Species

In this paper, the overlapping rectangular part of the two-source data was used as the study area. Using Lidar-360 software, the airborne point cloud data were used to construct CHM with an interval space of 0.5 m. A minimum tree height of 2 m, a buffer area of 50 pixels, a Gaussian smoothing factor of 0.7, and a radius of 5 pixels were used to segment the study forest into individual trees. By integrating the point cloud structure features and image index features of the single tree area, the sample data were used to classify the single tree species one by one. The sample test accuracy of the classification experiment was better than 99.4%. The classification result is shown in Table 4 and the classification result diagram is shown in Figure 3.

Table 4. Statistical results of tree classification in the study area.

Specie	*CL*	*PY*	*DBF*	Total
Quantity	388	343	1759	2490

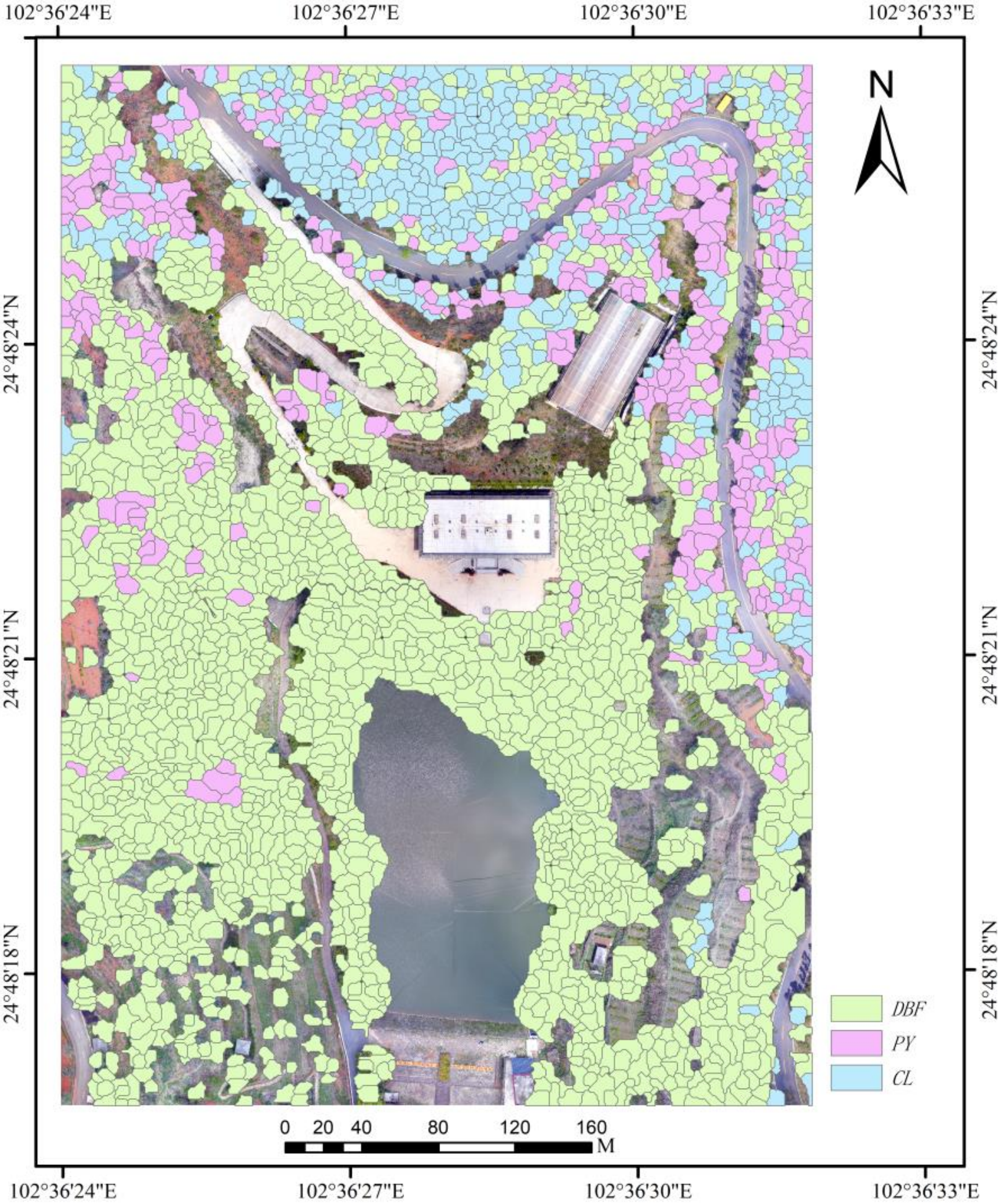

Figure 3. The classification result map of tree species.

2.4.3. Spectral Features Extraction of the Images

(1) Spectral and visible light vegetation index feature extraction

The visible light image of UAV reflects the reflection of trees to visible light through the three-color channels of R, G, and B. The visible light color channel contains little vegetation information. However, many previous studies have shown that some combinatorial operation of different color channels [14–16] can better reflect vegetation information. Therefore, in this paper, 25 commonly used visible light vegetation indexes and 3-color channels were selected to form 28 vegetation index variables. The selected vegetation index variables are shown in Table 5.

Table 5. Information of vegetation index variables.

Variable Name	Abbreviation	Variable Description	Reference
Red band	R	R = R	—
Green band	G	G = G	—
Blue band	B	B = B	—
Normalized red band	r	r = R/(R + G + B)	[14]
Normalized green band	g	g = G/(R + G + B)	[14]
Normalized blue band	b	b = B/(R + G + B)	[14]
Red–green ratio index	RGRI	RGRI = r/g	[14]
Green–blue ratio index	GBRI	GBRI = g/b	[14]
Red–blue ratio index	RBRI	RBVI = r/b	[14]
Red–blue difference index	RBDI	RBDI = r − b	[14]
Red–blue add index	RBAI	RBAI = r + b	[14]
Green–blue difference index	GBDI	GBDI = g − b	[14]
Red–blue vegetation index	RBVI	RBVI = (r − b)/(r + b)	[14]
Modified red–green–blue vegetation index	MRGBVI	MRGBVI = (r − b − g)/(r + g)	[14]
Modified green–red vegetation index	MGRVI	MGRVI = $(g^2 - r^2)/(g^2 + r^2)$	[14]
Red–green–blue vegetation index	RGBVI	RGBVI = $(g^2 - b \times r)/(g^2 + b \times r)$	[14]
Green–red vegetation index	GRVI	GRVI = (g − r)/(g + r)	[14]
Green leaf add index	GLA	GLA = (2 × g − r + b)/(2 × g + r + b)	[14]
Green leaf index	GLI	GLI = (2 × g − r − b)/(2 × g + r + b)	[14]
Excess red index	ExR	ExR = 1.4 × r − g	[14]
Excess green index	ExG	ExG = 2 × g − r − b	[14]
Excess green minus excess red index	ExGR	ExGR = ExG-1.4 × r − g	[14]
Color index of vegetation	CIVE	CIVE = 0.441 × r − 0.881 × g + 0.3856 × b + 18.78745	[14]
Vegetation atmospheric resistance index	VARI	VARI = (g − r)/(g + r − b)	[14]
Warbeck index	WI	WI = (g − b)/(r − g)	[14]
Normalized difference index	NDI	NDI = (r − g)/(r + g + 0.01)	[15]
Normalized green–blue difference index	NGBVI	NGBDI = (g − b)/(g + b)	[15]
Vegetation index	VEG	VEG = $g/(r^a \times b^{1-a})$; (a = 0.667)	[16]

(2) Image texture feature extraction

The gray level co-occurrence matrix (GLCM) [36], generally recognized as mature and effective, was used for texture feature extraction. The mean variance, homogeneity, contrast, dissimilarity, entropy, angular second moment (ASM), and correlation were used as statistical features. Multi-scale and multi-direction texture features were extracted with an interval of 2 and the four symmetric directions of angle 0, 45, 90, and 135. The calculation formulas of the texture statistics of images used in this study are shown in Table 6.

Table 6. Calculation formulas of statistical information for the texture of images.

Variable Name	Abbreviation	Variable Description	Reference
Mean	Mean	$\mathrm{Mean} = \sum_i \sum_j p_{(i,j)} * i$	[36]
Variance	Var	$\mathrm{Variance} = \sum_i \sum_j p_{(i,j)} * (i - \mathrm{Mean})^2$	
Homogeneity	Homo	$\mathrm{Homogeneity} = \sum_i \sum_j p_{(i,j)} * \frac{1}{1+(i-j)^2}$	
Contrast	Con	$\mathrm{Contrast} = \sum_i \sum_j p_{(i,j)} * (i - j)^2$	
Dissimilarity	Dis	$\mathrm{Dissimilarity} = \sum_i \sum_j p_{(i,j)} * \lvert i - j \rvert$	
Entropy	Ent	$\mathrm{Entropy} = \sum_i \sum_j p_{(i,j)} * \ln(i, j)$	
ASM	ASM	$\mathrm{ASM} = \sum_i \sum_j {p_{(i,j)}}^2$	
Correlation	Cor	$\mathrm{Correlation} = \sum_i \sum_j ((i - \mathrm{Mean}) * (j - \mathrm{Mean}) * {p_{(i,j)}}^2 / \mathrm{Variance})$	

2.4.4. Structural Features Extraction of Point Cloud

Because the airborne Lidar point cloud can truly and intuitively reflect the height, density, and vertical structure information of trees, it greatly compensates for the lack of image data in the vertical structure features of trees. In this paper, Lidar-360 software was used to extract the 101-dimensional point cloud structural features, including tree height features, density features, intensity features, canopy cover, etc. These descriptions of features are shown in Tables 7–9. They showed the height correlation features of trees, the intensity correlation variables, and the density correlation characteristics of point clouds and other information, respectively.

Table 7. Point cloud height feature information of trees.

Variable Name	Abbreviation	Quantity	Variable Description
Average absolute deviation	E-ADD	1	$V = \frac{\sum_{i=1}^{n} \lvert Z_i - \overline{Z} \rvert}{n}$
Canopy relief rate	E-CRR	1	$V = \frac{Z_{mean} - Z_{min}}{Z_{max} - Z_{min}}$
Accumulate height percentiles	E-AIH	15	$AIH_{X\%} = \sum_{i=0}^{X\%} Z_i$(X = 1,5,10,20,25,30,40,50,60,70,75,80,90,95,99)
Interquartile range of accumulate height percentile	E-AIH_IQ	1	$AIH\_IQ = AIH_{75\%} - AIH_{25\%}$
Variable coefficient	E-CV	1	$V = \frac{Z_{std}}{Z_{mean}} \times 100\%$
Kurtosis	E-Kurtosis	1	$Kurtosis = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^4}{\sigma^4}$
Median of median absolute deviation	E-Mad median	1	The median absolute deviation of the median height value at all points in the region.
Maximum, minimum, mean, median, skewness, standard deviation, and variance	E-max, E-min, E-mean, E-median, E-skewness, E-stan, and E-var	7	The maximum, minimum, mean, median, skewness, standard deviation, and variance of all point heights in the region.
Quadratic power mean	E-SMS	1	$V = \sqrt{\frac{\sum_{i=1}^{n} Z_i^2}{n}}$
The mean to the third power	E-CMC	1	$V = \sqrt[3]{\frac{\sum_{i=1}^{n} Z_i^3}{n}}$
Percentile of height	E-P	15	$Elev = Z_{X\%}$ (X = 1,5,10,20,25,30,40,50,60,70,75,80,90,95,99)
Interquartile range of Percentile of height	E-PIQ	1	$V = Elev_{75\%} - Elev_{25\%}$

Table 8. Point cloud intensity feature information of trees.

Variable Name	Abbreviation	Quantity	Variable Description
Mean absolute deviation	I-ADD	1	$V = \frac{\sum_{i=1}^{n} \lvert I_i - \overline{I} \rvert}{n}$
Accumulate intensity percentiles	I-AII	15	$AII_{X\%} = \sum_{i=0}^{X\%} I_i$(X = 1,5,10,20,25,30,40,50,60,70,75,80,90,95,99)
Variable coefficient	I-CV	1	$V = \frac{I_{std}}{I_{mean}} \times 100\%$
Kurtosis	I-Kurtosis	1	$Kurtosis = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(I_i - \overline{I})^4}{\sigma^4}$
Median of median absolute deviation	I-Mad median	1	The median absolute deviation of the median intensity value at all points in the region.
Maximum, minimum, mean, median, skewness, standard deviation, and variance	I-max, I-min, I-mean, I-median, I-skewness, I-stan, and I-var	7	The maximum, minimum, mean, median, skewness, standard deviation, and variance of intensity values of all points in the region.
Percentile of intensity	I-P	15	$Int = I_{X\%}$ (X = 1, 5, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 95, and 99)
Interquartile range of percentile of intensity	I-PIQ	1	$V = Int_{75\%} - Int_{25\%}$

Table 9. Point cloud density and other feature information of trees.

Variable Name	Abbreviation	Quantity	Variable Description
Density variable	D-M	10	In the region, the point cloud data are divided into ten equal height slices from low to high and the proportion of echo numbers in each layer is the corresponding density variable.
Canopy cover	CC	1	$\mathrm{CC} = \frac{\mathrm{n_{veg}}}{\mathrm{n_{total}}}$ ($\mathrm{n_{veg}}$ is the number of vegetation points, $\mathrm{n_{total}}$ is the total number of points)
Leaf area index	LAI	1	$\mathrm{LAI} = \frac{\cos(\mathrm{ang}) \times \ln(\mathrm{GF})}{\mathrm{k}}$ (ang is the average scan Angle, GF is the gap rate, and k is the extinction coefficient)
Gap Fraction	GF	1	$\mathrm{GF} = \frac{\mathrm{n_{ground}}}{\mathrm{n}}$ ($\mathrm{n_{ground}}$ is the number of ground points whose height is lower than the height threshold and n is the total number of points)

2.4.5. Canopy and Topographic Features Extraction

In addition to the corresponding height, intensity, and density of the point cloud data, the characteristics of the crown diameter, area, volume, and tree height associated with biomass can be extracted after single-tree segmentation of the point cloud data. Besides, different topographic distributions have non-negligible effects on the trend of water pooling, soil surface water content, and accumulation of soil surface humus. Therefore, 8 main topographic factors, including elevation, slope, aspect, slope length, slope variability, aspect variability, topographic relief, and ground roughness, were selected to measure the impact of topography characteristics on biomass estimation. These features used for biomass estimation are shown in Table 10.

Table 10. The information of canopy and topography features.

Variable Name	Abbreviation	Variable Description
Elevation	H	H = H
Slope	Slope	slope = slope
Aspect	Aspect	aspect = aspect
Slope Length	SL	$\mathrm{SL} = \mathrm{DEM}/\sin(\frac{\mathrm{slope}*\pi}{180})$
Slope Variability	SOS	SOS = slope of slope
Aspect Variability	SOA	SOA = slope of aspect
Terrain fluctuates degree	TFD	TFD = maxDEM − minDEM (Neighborhood range = 12)
Terrain Roughness	TR	$\mathrm{TR} = 1/\cos(\frac{\mathrm{slope}*\pi}{180})$
Crown Diameter	CD	The average diameter of the projected region of crown amplitude.
Tree Height	TreeH	The height difference between the top of the tree and the ground.
Crown Area	CA	The area of the projected region of crown amplitude
Crown Volume	CV	The volume of the canopy of a tree.

2.5. Statistical Analysis

2.5.1. Correlation Analysis

Correlation analysis is necessary to screen the feature variables and eliminate redundant feature dimensions with little contributions, in viewing the possibility of redundancy between homologous and heterologous data features as well as the contribution of features to the model decision. However, the traditional Pearson and Spearman correlation coefficient analyses require high-quality data and have the limitation of mainly solving linear correlation problems. Thus, this study introduced Kendall Rank correlation coefficient, which can measure the correlation of nonlinear data. It can determine correlation through the ranking consistency between two feature vectors and has better robustness compared to the commonly used Pearson and Spearman correlation coefficients [37]. The calculation formula is as follows:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{1}$$

Among them, n_c is the logarithm of vectors with the same ordering, n_d is the logarithm of vectors with different ordering, and n is the sample size in total. The value range of the Kendall Rank correlation coefficient is between (–1 and 1). The larger the coefficient absolute value, the higher the correlation is. The result (0.8, 1) indicates a very strong correlation, (0.6, 0.8) indicates a strong correlation, (0.4, 6) indicates a moderate correlation, (0.2, 0.4) indicates a weak correlation, and (0, 0.2) indicates a very weak correlation.

2.5.2. Permutation Importance Index Analysis

A machine learning algorithm based on a decision tree can determine the relative importance of each feature based on the measurement of impurity. However, such methods tend to amplify the importance of high-cardinality features and continuous features. Therefore, this study adopted the permutation importance (PI) index to measure feature importance, which was evaluated by observing the effect of the random rearrangement of each feature dimension on the model accuracy [38]. It has the advantages of convenient calculation, accurate feature evaluation, and good interpretability. The calculation steps are as follows.

Step 1: Train the model.

Step 2: Shuffle the columns of the feature data to be analyzed and analyze the importance of the feature vector by evaluating the change in accuracy.

Step 3: Restore the feature data and repeat Step 2 to analyze other feature vectors.

Because of the randomness of the single shuffled data, the experiment was repeated to evaluate the importance of the PI index generally.

2.6. Modeling

2.6.1. Random Forest

Random forest is a classical algorithm using a Bagging integration strategy. This algorithm integrates a large number of decision trees as the basic model. By random, put back a part of the data and features and the weak decision trees are integrated into a powerful model. Because this algorithm can deal with high-dimensional data and is not easy to overfit, it has been widely used in biomass estimation [31]. The key parameters of the algorithm are as follows: The number of decision trees (n_estimators): The greater the number of decision trees, the better the model effect will be and it will tend to be stable after reaching a certain number. Maximum tree depth (max_depth): It directly represents the complexity of the model. The maximum number of input features per tree (max_features): Increasing the number usually improves the model performance, but, at the same time, it reduces the tree diversity. In addition, the three key parameters are directly related to the complexity of the model. The larger the number, the more complex the model and the slower the processing speed.

2.6.2. XGBoost

Extreme gradient boosting (XGBoost) is an excellent algorithm improved on the basis of a gradient boosting decision tree (GBDT), in which, the regularization term is added to control its complexity and improve its generalization ability effectively [39]. Besides, this algorithm is widely used in many machine learning competitions after it is processed by efficient parallelization. Its key parameters include the number of decision trees (n_estimators), the maximum depth of the tree (max_depth), and the learning rate of the gradient descent (learning_rate). The learning rate can reduce the weight of each step and cause the model to be more robust. If the learning rate is too large, the accuracy will decrease, but if the value is too small, the model will run slowly.

2.6.3. Grid Search

Grid Search traverses all the parameter settings by setting the adjustment range of parameters and sampling method of exhaustion and uses the parameter combination with the best score as the best result. This method can fully exploit the estimation performance of each model and ensure the consistency of the experimental environment. Thus, it is widely used in various machine learning classifications and predictions.

2.6.4. Experimental Environment

The experimental environment of this study is AMD Ryzen-5 six-core CPU, GTX-1650 4 G GPU, 16 G memory, and Windows10 operating system. The software platform is MATLAB 2022a, Anaconda 3, and Python 3.7.

2.6.5. Model Evaluation Method

In order to compare the accuracy differences of different biomass estimation results, the root mean square error (RMSE) and R-Square coefficient of determination (R-square) were selected as the indexes to evaluate the estimation accuracy. The calculation formulas were as follows. In this paper, the ratio of 7:3 was adopted to randomly divide the training set and the test set. The accuracy and fitting performance of the model were verified by comparing the difference of evaluation indexes between the training set and the test set of different tree species.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{f_i} - \mathrm{y_i})^2} \tag{2}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{f_i} - \mathrm{y_i})^2}{\sum_{\mathrm{i}=1}^{\mathrm{m}}(\overline{\mathrm{y}}_\mathrm{i} - \mathrm{y_i})^2} \tag{3}$$

where, RMSE represents the deviation between the estimated result and the true value. The smaller the RMSE is, the smaller the deviation is. The numerator of R^2 represents the sum of the squared variance of the predicted value and the true value and the denominator represents the sum of the squared variance of the true value and the mean. Its value ranges from 0 to 1. A larger R^2 indicates a better estimation effect.

3. Results

3.1. Biomass Statistical Analysis of Tree Species Samples

The statistical information of the sample biomass calculated by the binary biomass equation is shown in Table 11. The results showed that the biomass distribution of the *Cupressus lusitanica* (*CL*) and the *Deciduous Broad-leaved Forest* (*DBF*) was uniform. However, the sample biomass of *Pinus yunnanensis* (*PY*) has a large difference with a variance of 64.533 kg and the largest weight span of the sample biomass was 273.294 kg.

Table 11. Biomass statistics of tree samples.

Class	Number	Minimum (kg)	Maximum (kg)	Mean (kg)	Variance (kg)
CL	36	5.086	81.560	26.597	19.735
PY	31	2.552	275.846	63.785	64.533
DBF	107	3.652	98.689	28.753	17.735

3.2. Feature Analysis

3.2.1. Correlation Analysis

By calculating the Kendall Rank coefficient between the multi-sourced features, we obtained the heat maps of the absolute value of the correlation coefficient, as shown in Figures 4–6. They clearly showed the degree of correlation among the features of different tree species. The X axis and the Y axis of the heat map both contain 173 dimensional-feature variables, a dependent variable, biomass, and a total of 174 calculation objects. The absolute value of the correlation between variables were rendered by gradients from blue to red. The closer the color is to red, the greater the correlation. The closer the color is to blue, the smaller the correlation. In addition, Figures 4–6 have separately enlarged the heat maps of feature correlation coefficients for objects 1–58, 59–116, and 117–174, which cause them to be easier to observe.

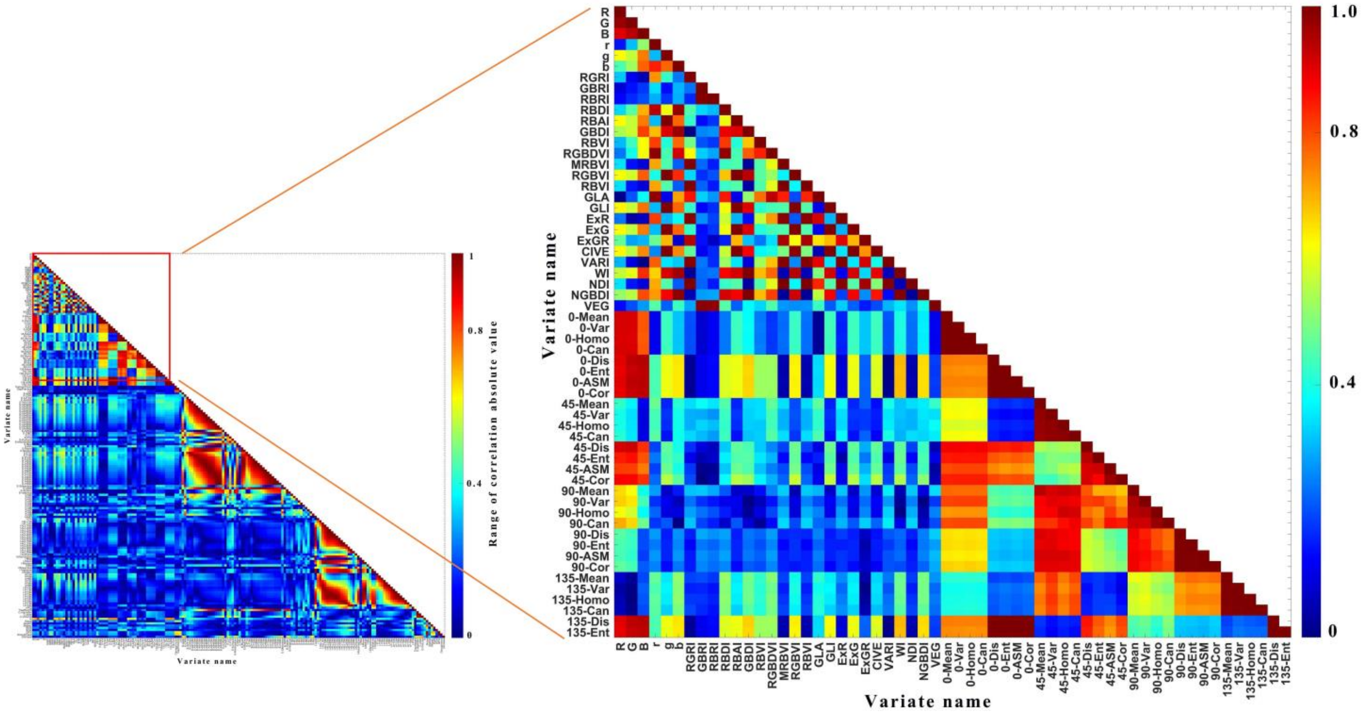

Figure 4. Heat map of absolute value of characteristic correlation coefficient of the *CL*.

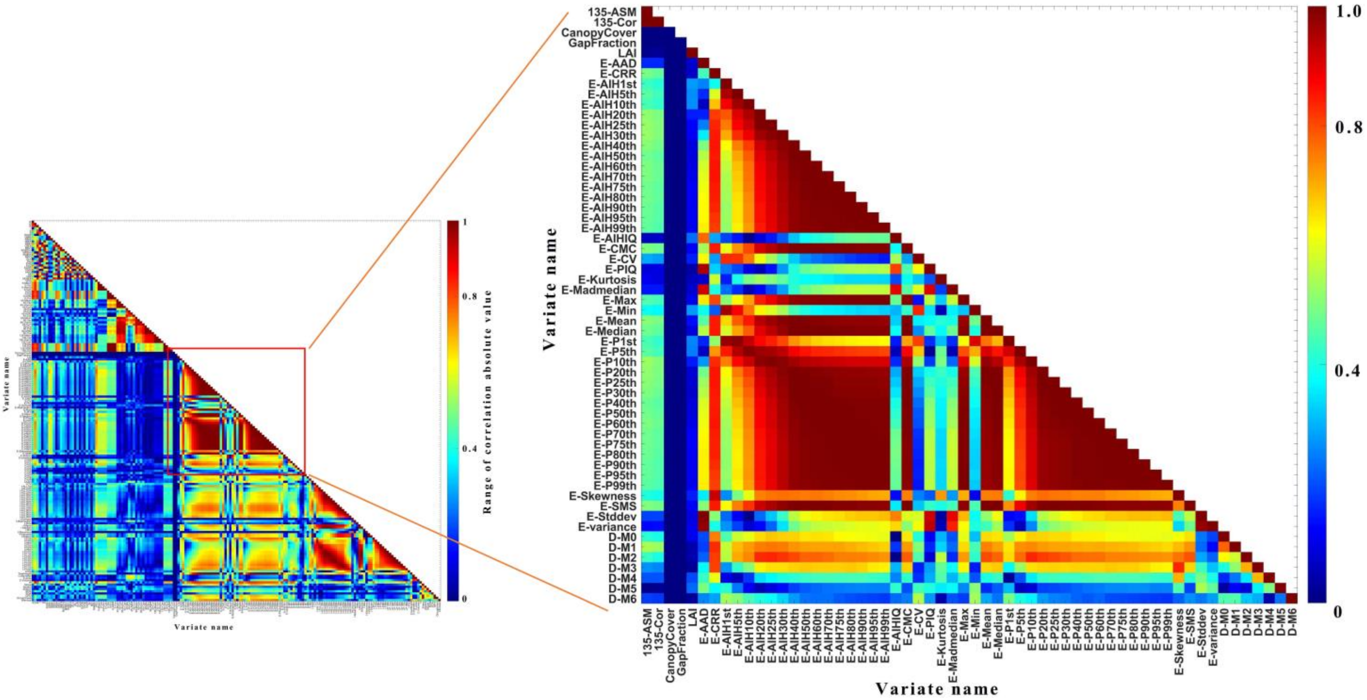

Figure 5. Heat map of absolute value of characteristic correlation coefficient of the *PY*.

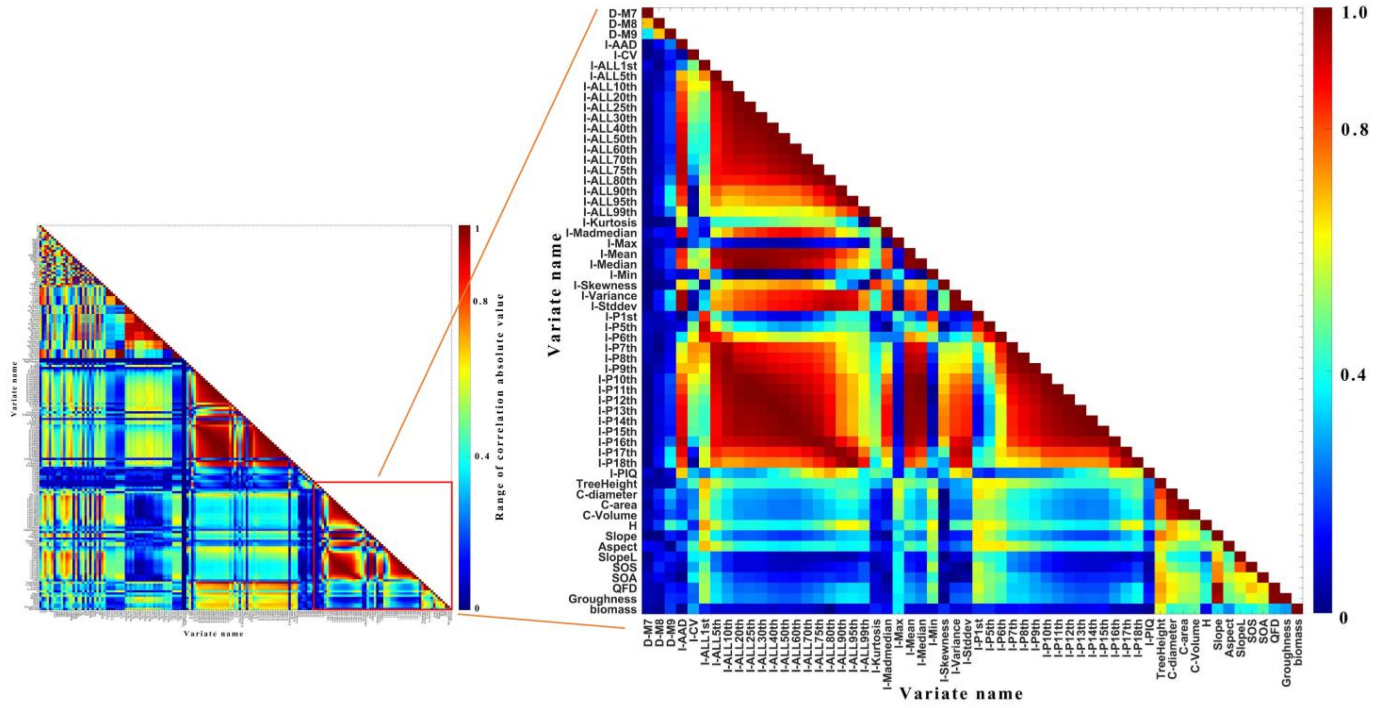

Figure 6. Heat map of absolute value of characteristic correlation coefficient of the *DBF*.

It was evident from the heat maps that the correlations of the traits between the different tree species were not the same. There were a large number of highly correlated feature pairs, most of which were clustered in the percentile and cumulative percentile features of the point cloud height and intensity features. In addition, the correlation between the homologous features was higher, whereas that between the heterologous features was lower.

The experiment showed that there were 411 pairs of features that were highly correlated in the *Cupressus lusitanica*, including 57 dimensional image indexes and texture features, 74 dimensional point cloud structural parameter features, and 6 dimensional canopy and topographic features. Additionally, 117 dimensional features were eliminated in the *Cupressus lusitanica*. There were 640 pairs of highly correlated features in the *Pinus yunnanensis*, involving a total of 132 dimensional features, i.e., 50 dimensional image features, 76 dimensional point cloud structural features, and 6 dimensional canopy and topographic features. Finally, 114 dimensional features were eliminated in the *Pinus yunnanensis* and, for the *Deciduous Broad-leaved Forest*, there were 758 pairs of features highly correlated, including 136 dimensional features, i.e., 57 dimensional image features, 72 dimensional point cloud features, and 7 other dimensional features. The 123 dimensional features were eliminated in the *Deciduous Broad-leaved Forest*. Table 12 shows the retained feature dimension information of each tree species after being filtered by the multicollinearity analysis.

Table 12. The feature information of each tree species were retained after multicollinearity analysis.

Tree Species	Quantity	Characteristics of Abbreviation
CL	56	R,G,B,r,g,RGRI,GBRI,VEG,0-Mean,0-Dis,45-Mean,45-Dis,90-Mean,90-Var,90-Dis,135-Mean,CanopyCover,LAI,E-AAD,E-CRR,E-AIH1st,E-AIHIQ,E-CV,E-PIQ,E-Kurtosis,E-Madmedian,E-Min,E-Skewness,D-M0,D-M1,D-M2,D-M3,D-M4,D-M5,D-M6,D-M7,D-M8,D-M9,I-AAD,I-CV,I-AII90th,I-Kurtosis,I-Madmedian,I-Max,I-Min,I-Skewness,I-PIQ,TreeHeight,C-Diameter,C-Volume,H,Slope,Aspect,Slopelength,SOS,SOA
PY	59	R,G,B,r,g,b,RGRI,GBRI,VARI,VEG,0-Mean,0-Dis,45-Mean,45-Dis,45-Ent,45-ASM,45-Cor,90-Var,90-Dis,135-Mean,CanopyCover,GapFraction,LAI,E-AAD,E-CRR,E-AIHIQ,E-CV,E-Kurtosis,D-M0,D-M1,D-M2,D-M3,D-M4,D-M5,D-M6,D-M7,D-M8,D-M9,I-AAD,I-CV,I-AII1st,I-AII99th,I-Kurtosis,I-Madmedian,I-Max,I-Min,I-Skewness,I-Variance,I-P1st,I-PIQ,TreeHeight,C-Diameter,C-Volume,H,Slope,Aspect,SlopeL,SOS,SOA
DBF	50	R,G,B,RGRI,GBRI,ExGR,0-Mean,0-Dis,45-Mean,90-Dis,135-Mean,CanopyCover,GapFraction,LAI,E-CRR,E-AIH1st,E-CV,E-Kurtosis,E-M,E-Skewness,D-M0,D-M1,D-M2,D-M3,D-M4,D-M5,D-M6,D-M7,D-M8,D-M9,I-AAD,I-CV,I-AII1st,I-AII99th,I-Kurtosis,I-Madmedian,I-Max,I-Min,I-Skewness,I-P1st,I-P10th,I-PIQ,TreeHeight,C-Diameter,H,Slope,Aspect,Slopelength,SOS,SOA

In addition to the multicollinearity analysis of the variables, the correlation analysis between the independent variables and biomass can reflect the importance of characteristics. The correlation of each tree species between the biomass and variables is shown in Figures 7–9, which reflect the rankings of the top 32 correlation absolute values of each tree species. The absolute value range of correlation is [0–1]; a larger value indicates a stronger correlation. In addition, the threshold of correlation for absolute value is set as 0.1. When the absolute value of the correlation coefficient between a variable and biomass is lower than 0.1, it needs to be eliminated as it indicates no correlation. Without additional multicollinearity analysis, the feature information can be retained after correlation analysis, as shown in Table 13.

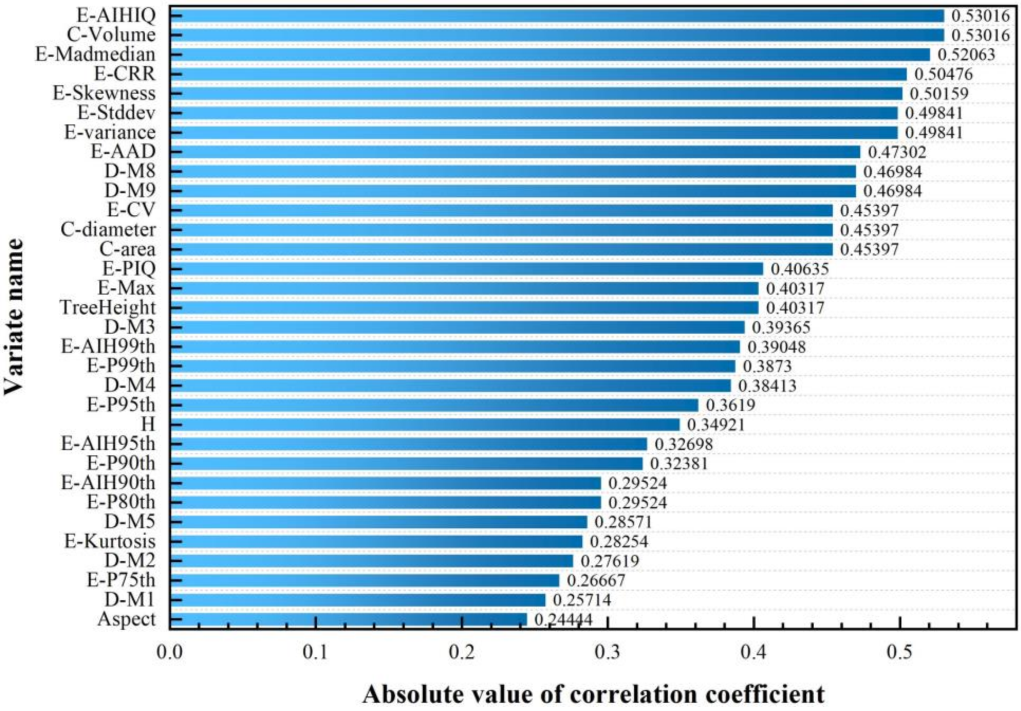

Figure 7. Ranking the 1–32 correlation characteristics of the *CL*.

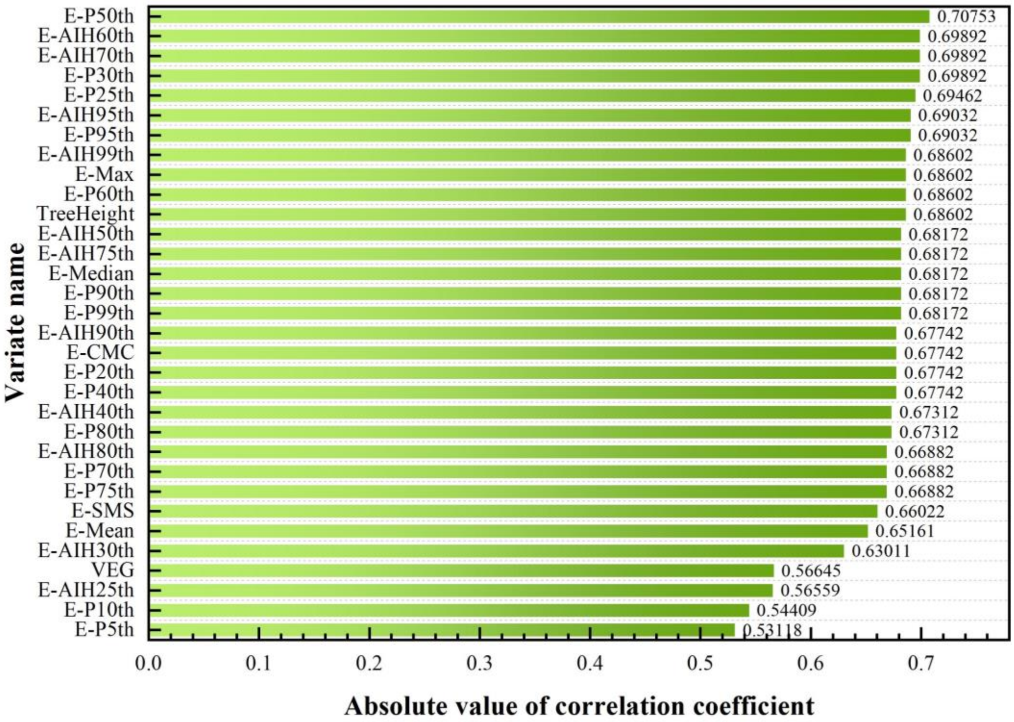

Figure 8. Ranking the 1–32 correlation characteristics of the *PY*.

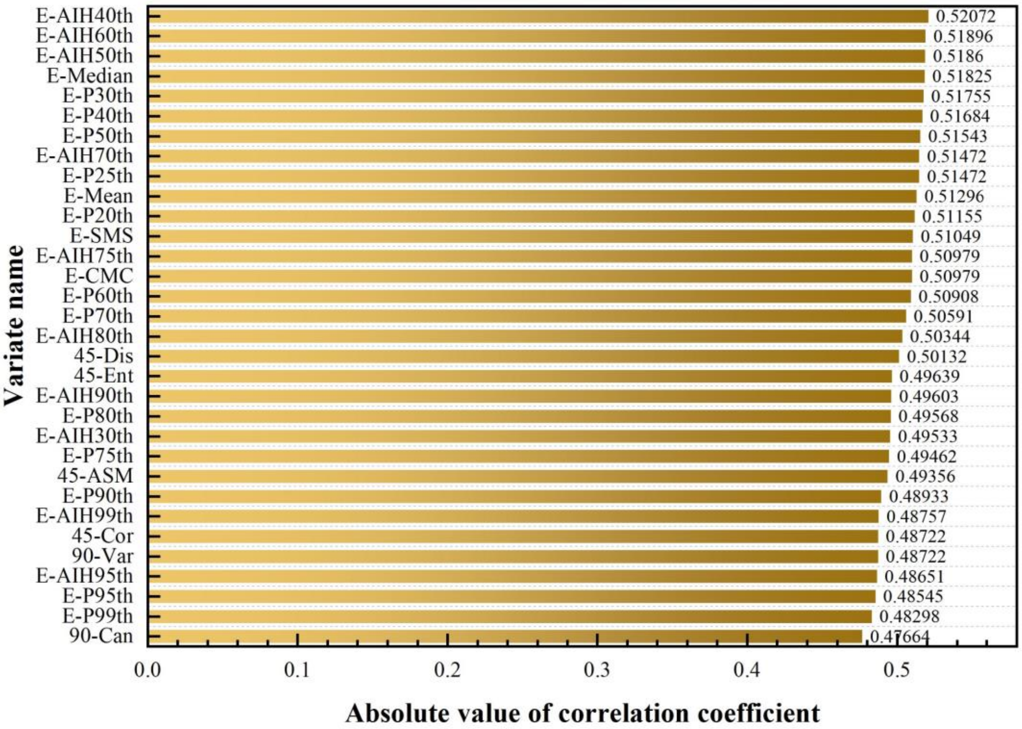

Figure 9. Ranking the 1–32 correlation characteristics of the *DBF*.

Table 13. The characteristic information retained after filtering the correlation of each tree species variable.

Tree Species	Quantity	Characteristics of Abbreviation
CL	70	r,RBDI,GBDI,RBVI,RGBDVI,WI,NGBDI,90-Dis,90-Ent,90-ASM,90-Cor,GapFraction,E-AAD,E-AIH70th,E-AIH75th,E-AIH80th,E-AIH90th,E-AIH95th,E-AIH99th,E-AIHIQ,E-CMC,E-CV,E-PIQ,E-Madmedian,E-Max,E-P40th,E-P50th,E-P60th,E-P70th,E-P75th,E-P80th,E-P90th,E-P95th,E-P99th,E-Skewness,E-SMS,E-Stddev,E-variance,D-M0,D-M1,D-M2,D-M3,D-M4,D-M5,D-M6,I-ALL1st,I-ALL5th,I-ALL10th,I-ALL20th,I-ALL25th,I-ALL30th,I-ALL40th,I-Max,I-Mean,I-Median,I-P1st,I-P8th,I-P9th,I-P10th,I-P11th,I-P12th,I-P13th,TreeHeight,C-diameter,C-area,C-Volume,Slope,Aspect,QFD,Groughness
PY	87	R,G,B,GBRI,RBRI,MRBVI,RBVI,GLA,ExGR,VARI,VEG,0-Mean,0-Var,0-Homo,0-Can,0-Dis,0-Ent,0-ASM,0-Cor,45-Mean,45-Var,45-Homo,135-Dis,135-Ent,135-ASM,135-Cor,CanopyCover,GapFraction,E-AAD,E-CRR,E-AIH1st,E-AIH5th,E-AIH10th,E-AIH20th,E-AIH25th,E-AIH30th,E-AIH40th,E-AIH50th,E-AIH60th,E-AIH70th,E-AIH75th,E-AIH80th,E-AIH90th,E-AIH95th,E-AIH99th,E-AIHIQ,E-CMC,E-PIQ,E-Kurtosis,E-Madmedian,E-Max,E-Min,E-Mean,E-Median,E-P1st,E-P5th,E-P10th,E-P20th,E-P25th,E-P30th,E-P40th,E-P50th,E-P60th,E-P70th,E-P75th,E-P80th,E-P90th,E-P95th,E-P99th,E-SMS,E-Stddev,E-variance,D-M6,D-M7,D-M8,D-M9,I-CV,TreeHeight,C-diameter,C-area,C-Volume,H,Slope,Aspect,SOS,QFD,Groughness
DBF	77	b,GBRI,RBRI,MRBVI,RBVI,GLA,ExGR,VARI,VEG,45-Mean,45-Var,45-Homo,45-Can,135-Mean,135-Var,135-Homo,135-Can,CanopyCover,GapFraction,E-AAD,E-CRR,E-AIH1st,E-AIH5th,E-AIH10th,E-AIH20th,E-AIH25th,E-AIH30th,E-AIH40th,E-AIH50th,E-AIH60th,E-AIH70th,E-AIH75th,E-AIH80th,E-AIH90th,E-AIH95th,E-AIH99th,E-AIHIQ,E-CMC,E-CV,E-PIQ,E-Madmedian,E-Max,E-Mean,E-Median,E-P1st,E-P5th,E-P10th,E-P20th,E-P25th,E-P30th,E-P40th,E-P50th,E-P60th,E-P70th,E-P75th,E-P80th,E-P90th,E-P95th,E-P99th,E-SMS,E-Stddev,E-variance,D-M6,D-M9,I-CV,I-ALL90th,I-Kurtosis,I-Max,I-Skewness,TreeHeight,C-diameter,C-area,C-Volume,Slope,SOS,QFD,Groughness

The analysis of the correlation between the feature variables and biomass showed great heterogeneity among the different tree species. A higher correlation was observed between the biomass of the *Cupressus lusitanica* and the variables of point cloud height and density, as well as the canopy and topographic features. However, the correlations of the vegetation index, texture, and point cloud intensity characteristics with biomass of this species were lower than the top 32. For the *Pinus yunnanensis*, the correlation of the point cloud height variable was the highest and the correlation between the vegetation index and tree height was also in the top 32. However, there was no observed correlation with the density variables, as in that observed among the *Cupressus lusitanica*. The *Deciduous Broad-leaved Forest* showed strong correlations with the image texture and point cloud height variables. In conclusion, the point cloud height related variables were generally strongly correlated with the biomass of different tree species. The features related to the point cloud, tree canopy, and topography were more preserved after screening. The visible light vegetation index and texture features related to the image lacked the correlation with biomass. In Table 13, more than half or even five-sixths of the image variables were excluded for each of the three tree species.

3.2.2. Permutation Importance Index of Features

In this section, RF is used as the base model for permutation importance (PI) analysis. The maximum depth of the decision trees was set to 5, the number of decision trees was set to 500, and the training set and the test set were divided by 7:3. The PI index was evaluated through ten repeated experiments to judge the importance of each characteristic of different tree species for biomass regression. The PI index results are shown in Figure 10. After screening, the mean value of the tenfold PI index was greater than 0, indicating a positive effect on the estimation model. The information on the features that need to be retained is shown in Table 14.

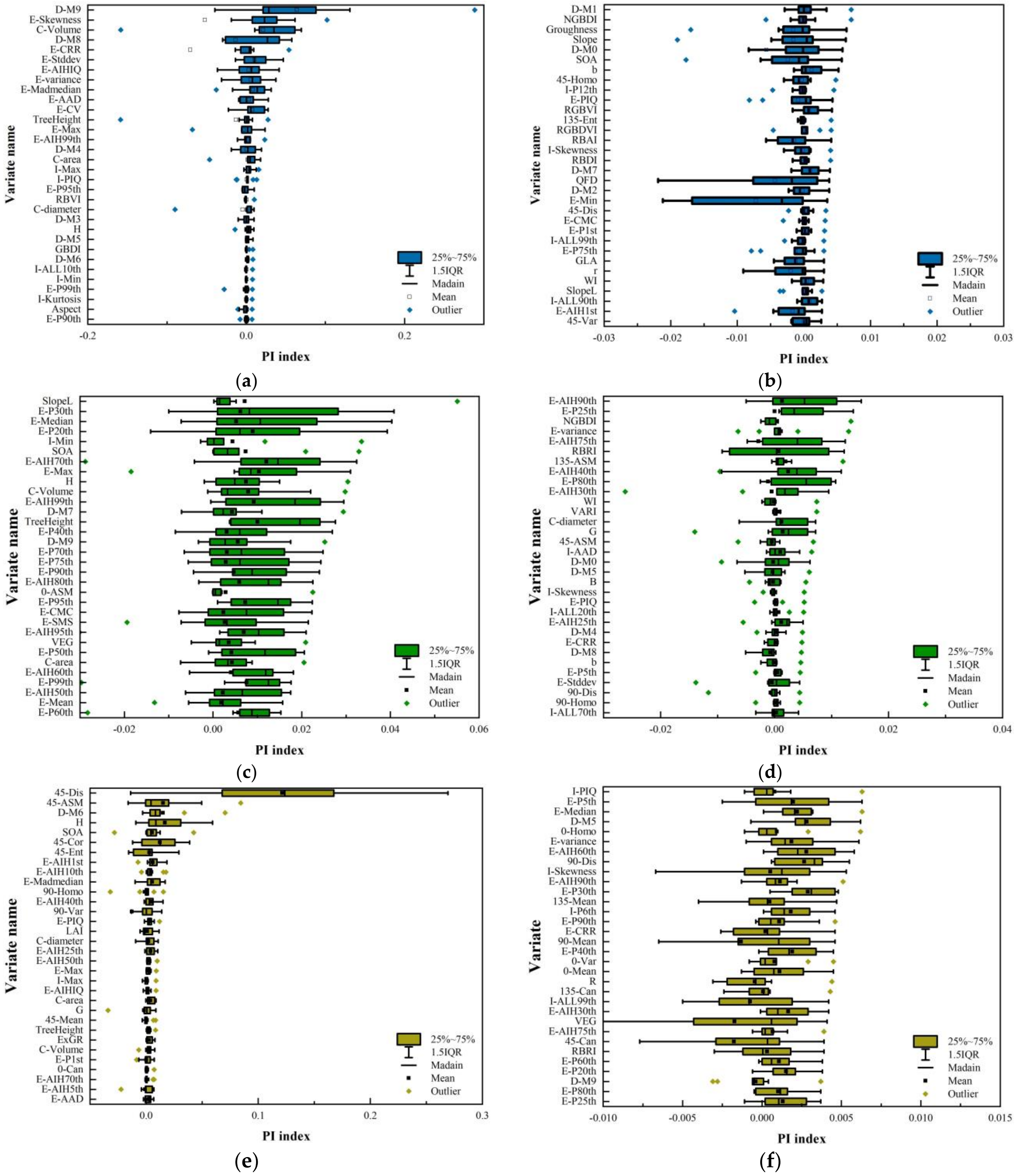

Figure 10. PI index ranking of some important variables. (**a**) The top 1–32 PI importance characteristics of the *CL*. (**b**) The top 33–64 PI importance characteristics of the *CL*. (**c**) The top 1–32 PI importance characteristics of the *PY*. (**d**) The top 33–64 PI importance characteristics of the *PY*. (**e**) The top 1–32 PI importance characteristics of the *DBF*. (**f**) The top 33–64 PI importance characteristics of the *DBF*.

Table 14. The feature information retained after filtering the variable PI index of each tree species.

Tree Species	Quantity	Characteristics of Abbreviation
CL	53	G,B,b,GBRI,RBDI,GBDI,RBVI,RGBDVI,RGBVI,GLI,ExG,WI,NGBDI,0-Dis,0-ASM,45-Var,45-Dis,45-ASM,90-Can,135-Can,135-Ent,135-ASM,E-AAD,E-AIH99th,E-AIHIQ,E-CMC,E-CV,E-Kurtosis,E-Madmedian,E-P1st,E-Stddev,E-variance,D-M1,D-M3,D-M4,D-M5,D-M6,D-M7,D-M9,I-ALL10th,I-ALL90th,I-ALL95th,I-Kurtosis,I-Madmedian,I-Max,I-Min,I-P1st,I-P13th,I-P18th,I-PIQ,C-area,C-Volume,H
PY	101	R,G,r,RGRI,RBRI,RBDI,RGBDVI,MRBVI,ExGR,VARI,NGBDI,VEG,0-Var,0-Homo,0-Dis,0-Ent,0-ASM,0-Cor,45-Ent,45-Cor,90-Homo,135-Dis,135-Ent,135-ASM,135-Cor,E-AAD,E-CRR,E-AIH1st,E-AIH5th,E-AIH10th,E-AIH20th,E-AIH25th,E-AIH40th,E-AIH50th,E-AIH60th,E-AIH70th,E-AIH80th,E-AIH90th,E-AIH95th,E-AIH99th,E-AIHIQ,E-CMC,E-PIQ,E-Madmedian,E-Max,E-Mean,E-Median,E-P5th,E-P20th,E-P25th,E-P30th,E-P40th,E-P50th,E-P60th,E-P70th,E-P75th,E-P90th,E-P95th,E-P99th,E-SMS,E-variance,D-M1,D-M2,D-M4,D-M7,D-M9,I-AAD,I-ALL1st,I-ALL5th,I-ALL10th,I-ALL20th,I-ALL25th,I-ALL50th,I-ALL70th,I-ALL99th,I-Madmedian,I-Min,I-Skewness,I-Variance,I-P5th,I-P6th,I-P7th,I-P8th,I-P9th,I-P11th,I-P12th,I-P13th,I-P14th,I-P16th,I-P17th,I-P18th,I-PIQ,TreeHeight,C-diameter,C-area,C-Volume,H,SlopeL,SOS,SOA,QFD
DBF	97	g,b,RBRI,RBAI,GBDI,MRBVI,RGBVI,ExGR,0-Mean,0-Var,0-Homo,0-Can,0-Dis,45-Mean,45-Dis,45-Ent,45-ASM,45-Cor,90-Dis,90-Ent,90-ASM,135-Mean,135-Can,135-Dis,135-Ent,135-ASM,135-Cor,LAI,E-AAD,E-CRR,E-AIH1st,E-AIH10th,E-AIH20th,E-AIH25th,E-AIH30th,E-AIH40th,E-AIH50th,E-AIH60th,E-AIH70th,E-AIH75th,E-AIH80th,E-AIH90th,E-AIH95th,E-AIH99th,E-AIHIQ,E-CMC,E-PIQ,E-Madmedian,E-Max,E-Mean,E-Median,E-P1st,E-P5th,E-P10th,E-P20th,E-P25th,E-P30th,E-P40th,E-P50th,E-P60th,E-P70th,E-P75th,E-P80th,E-P90th,E-P95th,E-P99th,E-SMS,E-Stddev,E-variance,D-M0,D-M2,D-M5,D-M6,I-ALL1st,I-ALL25th,I-ALL60th,I-ALL80th,I-ALL90th,I-Max,I-Mean,I-Min,I-Skewness,I-P5th,I-P6th,I-P8th,I-P10th,I-P11th,I-P13th,I-P15th,I-P16th,I-PIQ,TreeHeight,C-diameter,C-area,C-Volume,H,SOA

When analyzing the characteristic PI index of the three tree species, the variables related to point cloud height, canopy, and topographic features were generally more important in estimating the biomass, while the visible light index of the image was relatively lower. The PI importance scores of the point cloud height percentile, cumulative percentile, and its height statistics, canopy volume, tree height, and ground elevation as individual variables were significant. It was found that the density and intensity features of the point clouds were more important for the *CL* than for the other two tree species and the image texture was more important only for the DBF.

3.3. Range of Grid Search Parameters

In order to eliminate the influence of parameter settings on the comparison of the biomass estimation with different characteristic combinations, Grid Search was used to investigate the estimation performance of each model. The consistency of the experimental parameter environment was maintained during the Grid Search. The search ranges of the parameters were shown in Table 15.

Table 15. Search ranges of grid parameters of each model.

Model Name	Parameters Range
RF	Number of decision trees (n_estimators): [50,300], Sampling interval: 20 Maximum depth of the model (max_depth): [3,10], Sampling interval: 2 Maximum number of input features in a single tree (max_features): [2,20], Sampling interval: 2
XGBoost	Number of decision trees (n_estimators): [50,300], Sampling interval: 20 Maximum depth of the model (max_depth): [3,10], Sampling interval: 2 Learning rate of gradient descent (learning_rate): [0.01,0.3], Sampling interval: 0.01

3.4. Model Construction and Evaluation using Original Feature Combinations

In this section, the RF and XGBoost models for biomass estimation of the three dominant tree species were constructed based on the different combinations of the features. The accuracy evaluation results are shown in Tables 16 and 17. The results showed that XGBoost had a stronger learning ability than the RF model, which resulted in a large difference in the accuracy between the training set and the test set. There was a serious problem of overfitting. The estimation results of different tree species indicated that the point cloud structural features were more suitable for biomass estimation, compared with the image features, because they achieved higher estimation accuracy and a better fitting effect. In addition, the estimation accuracy of the model biomass could be significantly improved by combining different types of features. The optimal or suboptimal test set accuracy was obtained by combining all the features in these comparative experiments.

Table 16. Comparison of regression accuracy of RF model with different original feature combinations.

Species	Class	Number	Max Depth	Estimators	Max Features	RMSE of Train Set	R^2 of Train Set	RMSE of Test Set	R^2 of Test Set
CL	Image	60	9	170	2	0.1157	0.8163	0.2061	0.1859
	Point cloud	101	3	90	4	0.1150	0.8186	0.1315	0.6684
	Image + Point cloud	161	3	50	16	0.1127	0.8258	0.1334	0.6586
	Image + Other	72	9	130	14	0.1008	0.8606	0.1850	0.3440
	Point cloud + Other	113	9	110	6	**0.0951**	**0.8760**	0.1219	0.7152
	ALL	173	5	130	12	0.0999	0.8632	**0.1209**	**0.7197**
PY	Image	60	9	50	2	0.0824	0.8679	0.2056	0.3479
	Point cloud	101	5	90	18	0.0467	0.9576	0.1123	0.8055
	Image + Point cloud	161	9	50	14	0.0463	0.9583	0.1087	0.8176
	Image + Other	72	9	50	16	0.0643	0.9194	0.1748	0.5285
	Point cloud + Other	113	7	50	6	0.0463	0.9582	0.1119	0.8068
	ALL	173	9	50	18	**0.0347**	**0.9766**	**0.1035**	**0.8347**
DBF	Image	60	5	50	12	0.0660	0.8803	0.1158	0.5622
	Point cloud	101	9	90	12	**0.0501**	**0.9309**	0.1107	0.5996
	Image + Point cloud	161	5	50	8	0.0635	0.8890	0.1049	0.6406
	Image + Other	72	5	50	16	0.0645	0.8855	**0.1034**	**0.6508**
	Point cloud + Other	113	9	50	12	0.0574	0.9092	0.1146	0.5706
	ALL	173	9	110	16	0.0542	0.9193	0.1045	0.6429

Table 17. Comparison of regression accuracy of XGBoost model with different original feature combinations.

Species	Class	Number	Max Depth	Estimators	Max Features	RMSE of Train Set	R^2 of Train Set	RMSE of Test Set	R^2 of Test Set
CL	Image	60	7	50	0.55	**0.0003**	**1.0000**	0.3062	−0.7972
	Point cloud	101	3	210	0.03	0.1183	0.8081	0.1429	0.6084
	Image + Point cloud	161	7	50	0.43	**0.0003**	**1.0000**	0.2030	0.2102
	Image + Other	72	3	50	0.53	**0.0003**	**1.0000**	0.1943	0.2766
	Point cloud + Other	113	3	50	0.13	0.1018	0.8580	**0.1330**	**0.6611**
	ALL	173	3	110	0.16	0.0021	0.9999	0.1175	0.5489
PY	Image	60	3	50	0.47	0.0003	1.0000	0.2268	0.2070
	Point cloud	101	3	50	0.57	0.0003	1.0000	0.2154	0.2845
	Image + Point cloud	161	3	90	0.21	0.0004	1.0000	0.1985	0.3923
	Image + Other	72	3	50	0.69	**0.0002**	**1.0000**	**0.1320**	**0.7314**
	Point cloud + Other	113	5	70	0.43	**0.0002**	**1.0000**	0.1982	0.3943
	ALL	173	5	70	0.43	0.0003	1.0000	0.1765	0.5195
DBF	Image	60	5	90	0.25	0.0003	1.0000	0.1191	0.5363
	Point cloud	101	7	50	0.69	**0.0002**	**1.0000**	0.1150	0.5680
	Image + Point cloud	161	7	70	0.35	**0.0002**	**1.0000**	0.1115	0.5936
	Image + Other	72	5	50	0.47	0.0003	1.0000	0.1122	0.5888
	Point cloud + Other	113	9	50	0.45	**0.0002**	**1.0000**	**0.1091**	**0.6108**
	ALL	173	3	50	0.47	**0.0002**	**1.0000**	0.1103	0.6025

3.5. Model Construction and Evaluation Using Filtered Features

Due to the multicollinearity, weak correlation, and low contribution of the features, the experiment in this section evaluated the effect of the feature selection methods on the accuracy of biomass regression by comparing three feature selection methods: multicollinearity analysis, variable correlation filtrating, and permutation importance index. The results are shown in Tables 18 and 19. After the multicollinearity analysis (MA), the test accuracy of the model is significantly reduced compared to all the features trained in the previous section. The overfitting problem of the XGBoost model is still serious. The multicollinearity analysis did not improve the accuracy of the RF and XGBoost models. The test accuracy of the PI feature importance screening method was better than that of the correlation coefficient method and the RF model combined with the PI filtering method achieved the best test set accuracy in the three species.

Table 18. Comparison of regression accuracy of RF with different feature selecting methods.

Species	Method	Number	Max Depth	Estimators	Max Features	RMSE of Train Set	R^2 of Train Set	RMSE of Test Set	R^2 of Test Set
CL	MA	56	3	50	4	0.1005	0.8614	0.1248	0.7013
PY		59	7	50	18	0.0492	0.9528	0.1605	0.6029
DBF		50	9	70	18	0.0565	0.9122	0.1023	0.6578
CL	CC	70	7	150	8	**0.0976**	**0.8695**	0.1188	0.7296
PY		87	7	70	10	0.0518	0.9477	0.1058	0.8275
DBF		77	9	270	18	**0.0515**	**0.9272**	0.1162	0.5589
CL	PI	53	3	70	4	0.1225	0.7942	**0.1176**	**0.7346**
PY		101	9	50	16	**0.0385**	**0.9712**	**0.0960**	**0.8578**
DBF		97	5	70	16	0.0584	0.9061	**0.0986**	**0.6823**

Table 19. Comparison of regression accuracy of XGBoost with different feature selecting methods.

Species	Method	Number	Max Depth	Estimators	Max Features	RMSE of Train Set	R^2 of Train Set	RMSE of Test Set	R^2 of Test Set
CL	MA	56	3	90	0.07	0.1166	0.8137	0.1653	0.4763
PY		59	3	50	0.47	**0.0003**	**1.0000**	0.1532	0.6382
DBF		50	5	50	0.63	0.0002	1.0000	0.1112	0.5960
CL	CC	70	7	50	0.13	0.1022	0.8567	0.1694	0.4496
PY		87	3	70	0.43	0.0003	1.0000	0.1677	0.5664
DBF		77	9	50	0.69	**0.0002**	**1.0000**	**0.1018**	**0.6614**
CL	PI	53	3	70	0.39	**0.0003**	**1.0000**	**0.1439**	**0.6032**
PY		101	7	50	0.55	0.0003	1.0000	**0.1319**	**0.7318**
DBF		97	3	50	0.55	0.0002	1.0000	0.1193	0.5349

3.6. Mapping of Biomass in The Study Area

Based on the optimal test results of the experiment in Section 3.5, the PI method was adopted to filter the retained feature dimension information and the trained model was used to estimate the biomass of the study area. The obtained biomass classification and grading diagram of the study area is shown in Figure 11 and the biomass statistical results are shown in Table 20. The analysis results showed that the *Deciduous Broad-leaved Forest* with the largest number of individual trees had the highest total biomass of 66,371.33 kg. The total biomass of the arbor forest was 104,102.76 kg in the study area. In addition, the *Pinus yunnanensis* had the most individuals with high biomasses and the highest average biomass. Finally, the *Cupressus lusitanica* did not have individuals with high biomass and the biomass of its individual trees was mostly distributed between 23 and 45 kg.

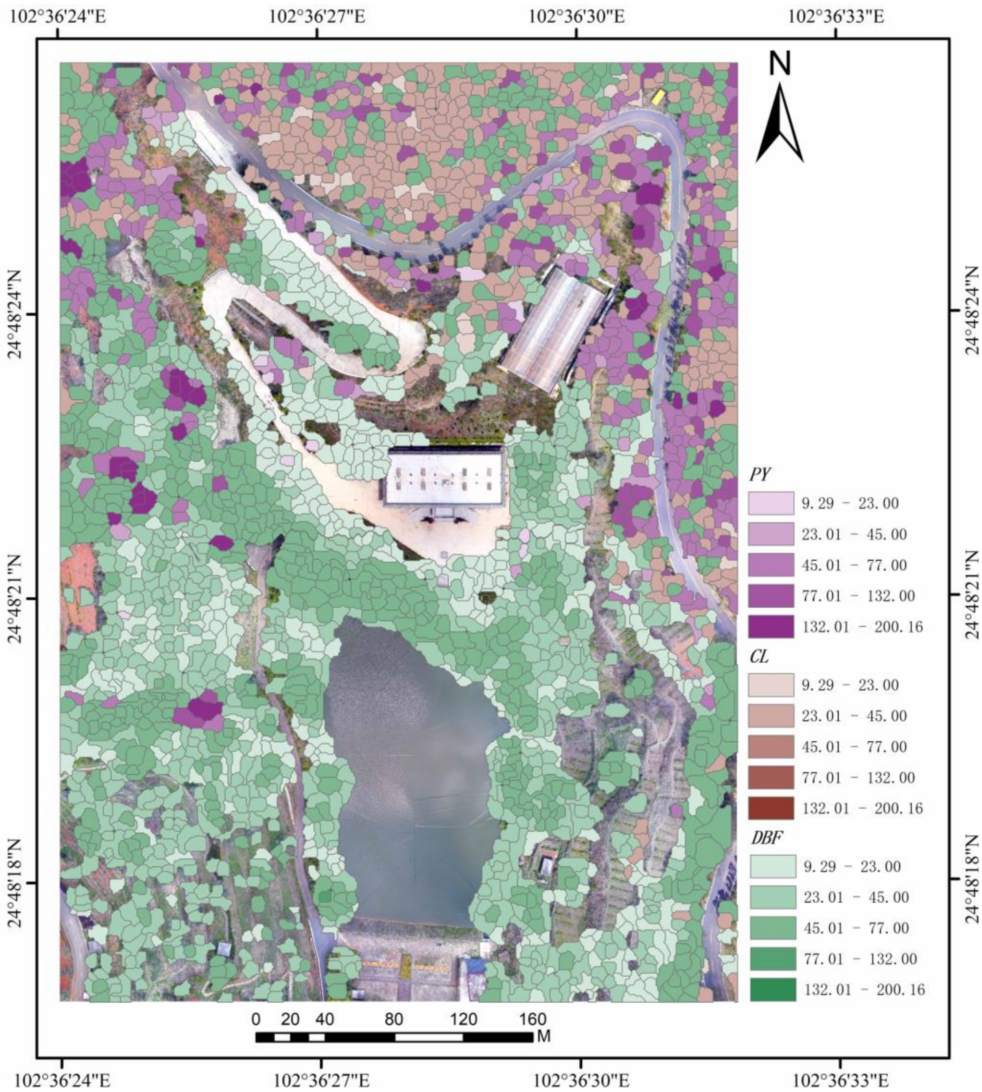

Figure 11. Forest biomass distribution at single tree scale in the study area.

Table 20. Biomass statistics of trees in study area.

Species	*Cupressus lusitanica*	*Pinus yunnanensis*	*Deciduous Broad-Leaved Forest*	Total
Maximum (kg)	44.96	200.16	77.06	—
Minimum (kg)	15.87	14.58	9.29	—
Mean (kg)	31.96	73.85	37.73	—
Variance (kg)	5.98	37.09	18.20	—
Total (kg)	12,399.61	25,331.82	66,371.33	104,102.76

4. Discussion

In this study, the strong correlations and important characteristic variables of the biomasses of different tree species in a subtropical plantation in Kunming, Yunnan province, were analyzed using visible ortho images of UAV and airborne lidar point cloud data. Based on random forest and XGBoost, different variable sets were used to construct estimation models, respectively. This paper mainly focused on the following areas:

(1) There were significant differences in the correlation and important features among different tree species. The point cloud height, canopy, and topographic features were all of high importance in the three tree species and were basically consistent with some previous research conclusions [26,27,40]. Among them, the point cloud height

percentile, cumulative percentile, canopy volume, tree height, and ground elevation as individual variables showed outstanding PI importance scores. The difference is that the features of the point cloud density and intensity are only of particular importance for the *Cupressus lusitanica* and the image texture features are only important for *Deciduous Broad-leaved Forest*.

Besides, our study result showed that the combinations of different categories of variables significantly improved the estimation accuracy and achieved better results than single point cloud feature estimation. In particular, the combination of point clouds, canopies, and topographic features greatly improved the biomass estimation accuracy of *Cupressus lusitanica* and *Pinus yunnanensis*. The image features have a great influence on the estimation precision of *Deciduous Broad-leaved Forests*, which further verified the conclusion of the PI index analysis.

(2) The estimation accuracy of RF and XGBoost models cannot be improved by eliminating the multicollinearity problem among variables. This feature-selection processing method has shown to be beneficial in the biomass estimation of linear models in the past [41,42]. However, for machine learning models that handle high-dimensional data well, such as RF and XGBoost, the sampling of features and samples can be adopted to avoid multicollinearity problems. At the same time, this processing will greatly reduce the diversity of the decision trees, which can affect the estimation accuracy. In addition, the estimation accuracy of the permutation importance index method is better than that of the correlation analysis. This indicates the advantages of wrapper methods such as PI in feature screening. Compared with the filter method, it is more targeted and has a better effect on model improvement [43].

(3) Image features such as the visible light vegetation index and texture had little influence on the biomass estimation of arboreal forests. Except for the texture features of the *Deciduous Broad-leaved Forest*, the image features were less correlated and important to the biomasses of the three tree species. In the feature combination experiment, the estimation accuracy was not improved much by combining the image features. This is different from the previous biomass estimation studies on annual crops such as maize, potato, and winter oilseed rape [19–21]. Because the spectral responses of annual crops are different in different growth stages and are associated with the accumulation process of crop biomass. The relatively accurate biomass estimation can be obtained through spectral information. Obviously, this theory is not applicable to the biomass estimation of subtropical arbor forests, whose leaf spectral responses are usually strongly correlated with seasons, but not with the growth cycle.

(4) The differentiation between species is beneficial for forest biomass estimation. The best estimation accuracy was achieved for all three tree species by using the combination of the post-fusion features, permutation importance index, and random forest model. The best estimation accuracy of the three tree species in the test sets are: *Cupressus lusitanica*: (RMSE = 0.1176, R^2 = 0.7346), *Pinus yunnanensis*: (RMSE = 0.0960, R^2 = 0.8578), and *Deciduous Broad-leaved Forest*: (RMSE = 0.0986, R^2 = 0.6823). Among them, the single species of *Cupressus lusitanica* and *Pinus yunnanensis* achieved significantly higher estimation accuracy than the mixed species of *Subtropical Deciduous Broad-leaved Forest*.

(5) The biomass estimation model in this study has regional limitations. Since the true values of the sample biomass were calculated by regional bivariate biomass equations, such equations are often strictly limited to the application area and tree species. Therefore, it is very important to change the acquisition method of the biomass truth value for reference in future research.

5. Conclusions

In this paper, multi-source remote sensing features were extracted based on UAV orthophotos and airborne Lidar point clouds and the features were filtrated using correlation analysis and a permutation importance index. The key features of the three dominant tree species are identified separately and the differences in their contributions to the estimation model were discussed. In addition, the biomass mapping of the single tree-scale forest in the study area was also completed through the fusion of image and point cloud features. The main results are as follows:

(1) In this study, the visible light vegetation index, texture, point cloud height, density, intensity, canopy, and topographic features were extracted. Among them, the point cloud height features, canopy variates, and the topographic-related factors showed great importance in the biomass estimation of the three species. In addition, the characteristics of the point cloud density and intensity are only important for *Cupressus lusitanica*. The image texture features are only important for the subtropical *Deciduous Broad-leaved Forest*. The visible light index had little effect on the biomass estimation of the three tree species.

(2) A total of 2490 individual trees were segmented in the study area, which included 388 strains of *Cupressus lusitanica*, 343 strains of *Pinus yunnanensis*, and 1759 strains of *Deciduous Broad-leaved Forest*. The total biomass of the forest in the study area was 104,102.76 kg, including 12,399.61 kg of *Cupressus lusitanica*, 25,331.82 kg of *Pinus yunnanensis*, and 66,371.33 kg of *Deciduous Broad-leaved Forest*. The mean biomass of *Pinus yunnanensis* was the highest (73.85 kg) and its number of individuals with high biomasses was the largest.

(3) From the estimation results of the two models, the random forest model with a poor learning ability showed better generalization ability in biomass estimation experiments with small samples and high dimensional data. Therefore, it is extremely necessary to collect more sample data when an estimation model with stronger learning ability is needed to improve the estimation accuracy.

Author Contributions: Conceptualization, M.Y. (Min Yan), Y.X., X.Y., D.W. and Y.H.; methodology, M.Y. (Min Yan), Y.X., X.Y., X.W., C.W., M.Y. (Minglong Yang), D.W. and Y.H.; software, M.Y. (Min Yan); validation, M.Y. (Min Yan), Y.X., X.Y. and C.W.; resources, Y.X., X.Y., X.W., C.W., M.Y. (Minglong Yang), D.W. and Y.H.; data curation, M.Y. (Min Yan), Y.X. and M.Y. (Minglong Yang); writing—original draft preparation, M.Y. (Min Yan), Y.X. and X.Y.; funding acquisition, Y.X., X.Y., X.W., C.W. and M.Y. (Minglong Yang). All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the National Natural Science Foundation of China (grant no. 41861054) and (grant no. 42161067).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the first author upon request.

Acknowledgments: First of all, we would like to thank Xiali Kong, Haoyu Tai, Chen Li, Yirong Pan, and Kunxiang Zhang for their help and suggestions during the data collection and paper revision. In addition, thanks to the *Lidar360* software trial code of *Beijing Digital Green Valley Technology Co., Ltd.* for the help with our data processing work. We also thank the reviewers and editors of *Sustainability* for their hard work.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Latella, M.; Raimondo, T.; Belcore, E.; Salerno, L.; Camporeale, C. On the integration of LiDAR and field data for riparian biomass estimation. *J. Environ. Manag.* **2022**, *322*, 116046. [CrossRef]
2. Qin, H.; Zhou, W.; Qian, Y.; Zhang, H.; Yao, Y. Estimating aboveground carbon stocks of urban trees by synergizing ICESat-2 LiDAR with GF-2 data. *Urban For. Urban Green.* **2022**, *76*, 127728. [CrossRef]
3. Luther, J.; Fournier, R.; Piercey, D.; Guindon, L.; Hall, R. Biomass mapping using forest type and structure derived from Landsat TM imagery. *Int. J. Appl. Earth Obs. Geoinform.* **2006**, *8*, 173–187. [CrossRef]

4. Hao, W.F.; Chen, C.; Liang, Z.; Ma, L. Research advances in vegetation biomass. *J. Northwest A F Univ. Nat. Sci. Ed.* **2008**, *36*, 175–182.
5. Zhang, C.; Peng, D.-L.; Huang, G.-S.; Zeng, W.-S. Developing Aboveground Biomass Equations Both Compatible with Tree Volume Equations and Additive Systems for Single-Trees in Poplar Plantations in Jiangsu Province, China. *Forests* **2016**, *7*, 32. [CrossRef]
6. Chakraborty, T.; Saha, S.; Reif, A. Biomass equations for European beech growing on dry sites. *iFor. Biogeosci. For.* **2016**, *9*, 751–757. [CrossRef]
7. Li, H.; Li, C.; Zha, T.; Liu, J.; Jia, X.; Wang, X.; Chen, W.; He, G. Patterns of biomass allocation in an age-sequence of secondary Pinus bungeana forests in China. *For. Chron.* **2014**, *90*, 169–176. [CrossRef]
8. Huang, G.; Li, Y. Phenological transition dictates the seasonal dynamics of ecosystem carbon exchange in a desert steppe. *J. Veg. Sci.* **2014**, *26*, 337–347. [CrossRef]
9. Cao, L.; Liu, H.; Fu, X.; Zhang, Z.; Shen, X.; Ruan, H. Comparison of UAV LiDAR and Digital Aerial Photogrammetry Point Clouds for Estimating Forest Structural Attributes in Subtropical Planted Forests. *Forests* **2019**, *10*, 145. [CrossRef]
10. Chen, S.; McDermid, G.J.; Castilla, G.; Linke, J. Measuring Vegetation Height in Linear Disturbances in the Boreal Forest with UAV Photogrammetry. *Remote Sens.* **2017**, *9*, 1257. [CrossRef]
11. Hasan, U.; Sawut, M.; Chen, S. Estimating the Leaf Area Index of Winter Wheat Based on Unmanned Aerial Vehicle RGB-Image Parameters. *Sustainability* **2019**, *11*, 6829. [CrossRef]
12. Wang, X.; Wang, Y.; Zhou, C.; Yin, L.; Feng, X. Urban forest monitoring based on multiple features at the single tree scale by UAV. *Urban For. Urban Green.* **2021**, *58*, 126958. [CrossRef]
13. Niu, Y.; Zhang, L.; Zhang, H.; Han, W.; Peng, X. Estimating Above-Ground Biomass of Maize Using Features Derived from UAV-Based RGB Imagery. *Remote Sens.* **2019**, *11*, 1261. [CrossRef]
14. Zheng, H.B.; Cheng, T.; Zhou, M.; Li, D.; Yao, X.; Tian, Y.C.; Cao, W.X.; Zhu, Y. Improved estimation of rice aboveground biomass combining textural and spectral analysis of UAV imagery. *Precis. Agric.* **2019**, *20*, 611–629. [CrossRef]
15. Wang, Z.; Ma, Y.; Chen, P.; Yang, Y.; Fu, H.; Yang, F.; Raza, M.; Guo, C.; Shu, C.; Sun, Y.; et al. Estimation of Rice Aboveground Biomass by Combining Canopy Spectral Reflectance and Un-manned Aerial Vehicle-Based Red Green Blue Imagery Data. *Front. Plant Sci.* **2022**, *13*, 903643. [CrossRef]
16. Liu, S.; Yang, G.; Jing, H.; Feng, H.; Li, H.; Chen, P.; Yang, W. Retrieval of winter wheat nitrogen content based on UAV digital image. *Trans. Chin. Soc. Agric. Engin.* **2019**, *35*, 75–85.
17. Liu, Y.; Huang, J.; Sun, Q.; Feng, H. Estimation of plant height and above ground biomass of potato based on UAV digital image. *Natl. Remote Sens. Bull.* **2021**, *25*, 2004–2014.
18. Che, Y.; Wang, Q.; Li, S.; Li, B.; Ma, Y. Monitoring of maize phenotypic traits using super-resolution reconstruction and multimodal data fusion. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 169–178.
19. Wang, L.; Zheng, N.; Chen, H.; Li, D.; Wu, M.; Zhao, W. Remote estimation of canopy height and aboveground biomass of maize using high-resolution stereo images from a low-cost unmanned aerial vehicle system. *Ecol. Indic.* **2016**, *67*, 637–648. [CrossRef]
20. Li, B.; Wu, X.; Zhang, L.; Han, J.; Bian, C.; Li, G.; Liu, J.; Jin, L. Above-ground biomass estimation and yield prediction in potato by using UAV-based RGB and hyperspectral imaging. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 161–172. [CrossRef]
21. Liu, Y.; Liu, S.; Li, J.; Guo, X.; Wang, S.; Lu, J. Estimating biomass of winter oilseed rape using vegetation indices and texture metrics derived from UAV multispectral images. *Comput. Electron. Agric.* **2019**, *166*, 105026. [CrossRef]
22. Li, C.; Yu, Z.; Wang, S.; Wu, F.; Wen, K.; Qi, J.; Huang, H. Crown Structure Metrics to Generalize Aboveground Biomass Estimation Model Using Airborne Laser Scanning Data in National Park of Hainan Tropical Rainforest, China. *Forests* **2022**, *13*, 1142. [CrossRef]
23. Liu, F.; Tan, C.; Lei, P.-F. Estimating individual tree aboveground biomass of the mid-subtropical forest using air-borne LiDAR technology. *J. Appl. Ecol.* **2014**, *25*, 3229–3236.
24. Rodríguez-Vivancos, A.; Manzanera, J.A.; Martín-Fernández, S.; García-Cimarras, A.; García-Abril, A. Analysis of structure from motion and airborne laser scanning features for the evaluation of forest structure. *Eur. J. For. Res.* **2022**, *141*, 447–465. [CrossRef]
25. Tao, S.; Guo, Q.; Li, L.; Xue, B.; Kelly, M.; Li, W.; Xu, G.; Su, Y. Airborne Lidar-derived volume metrics for aboveground biomass estimation: A com-parative assessment for conifer stands. *Agric. For. Meteorol.* **2014**, *198*, 24–32. [CrossRef]
26. Ullah, S.; Dees, M.; Datta, P.; Adler, P.; Koch, B. Comparing Airborne Laser Scanning, and Image-Based Point Clouds by Semi-Global Matching and Enhanced Automatic Terrain Extraction to Estimate Forest Timber Volume. *Forests* **2017**, *8*, 215. [CrossRef]
27. Gao, L.; Zhang, X. Above-Ground Biomass Estimation of Plantation with Complex Forest Stand Structure Using Multiple Features from Airborne Laser Scanning Point Cloud Data. *Forests* **2021**, *12*, 1713. [CrossRef]
28. Meng, Y.; Gou, R.; Bai, J.; Moreno-Mateos, D.; Davis, C.C.; Wan, L.; Song, S.; Zhang, H.; Zhu, X.; Lin, G. Spatial patterns and driving factors of carbon stocks in mangrove forests on Hainan Island, China. *Glob. Ecol. Biogeogr.* **2022**, *31*, 1692–1706. [CrossRef]
29. Wang, J.; Xiao, X.; Bajgain, R.; Starks, P.; Steiner, J.; Doughty, R.B.; Chang, Q. Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 189–201. [CrossRef]
30. Liu, Y.; Zhuang, Y.; Ji, B.; Zhang, G.; Rong, L.; Teng, G.; Wang, C. Prediction of laying hen house odor concentrations using machine learning models based on small sample data. *Comput. Electron. Agric.* **2022**, *195*, 106849. [CrossRef]

31. Hoover, C.M.; Ducey, M.J.; Colter, R.A.; Yamasaki, M. Evaluation of alternative approaches for landscape-scale biomass es-timation in a mixed-species northern forest. *For. Ecol. Manag.* **2018**, *409*, 552–563. [CrossRef]
32. Asner, G.P.; Mascaro, J.; Muller-Landau, H.; Vieilledent, G.; Vaudry, R.; Rasamoelina, M.; Hall, J.S.; van Breugel, M. A universal airborne LiDAR approach for tropical forest carbon mapping. *Oecologia* **2011**, *168*, 1147–1160. [CrossRef]
33. Zhang, Y.; Xia, C.; Zhang, X.; Cheng, X.; Feng, G.; Wang, Y.; Gao, Q. Estimating the maize biomass by crop height and narrowband vegetation indices derived from UAV-based hyperspectral images. *Ecol. Indic.* **2021**, *129*, 107985. [CrossRef]
34. Pan, P.; Li, R.; Xiang, C.; Zhu, Z.; Yin, X. Biomass and Productivity Of Cupressus Lusitanica Plantation. *Resour. Environ. Yangtze Val.* **2002**, *11*, 133–136.
35. Zhou, G.; Yin, G.; Tang, X.; Wen, D.; Liu, C.; Kuang, Y.; Wang, W. *Biomass Equation and Evaluation of Dominant Tree Species in China, Carbon Stocking-Biomass Equation of Forest Ecosystems in China*; Science Press: Beijing, China, 2018; pp. 40–80.
36. Hendrawan, Y.; Fauzi, M.R.; Khoirunnisa, N.S.; Andreane, M.; Hartianti, P.O.; Halim, T.D.; Umam, C. Development of colour co-occurrence matrix (CCM) texture analysis for biosensing. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *230*, 012022. [CrossRef]
37. Jin, L.; Li, Y. Analysis and realization of several correlation coefficients in R language. *J. Stat. Inf.* **2019**, *34*, 3–11.
38. Yu, H.; Xie, S.; Guo, L.; Liu, P.; Zhang, P. Extremely Randomized Trees Estimation of Soil Heavy Metal Content by Fusing Spectra and Spatial Features. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 231–239.
39. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *CoRR* **2016**, *1603*, 02754.
40. Tian, Y.; Huang, H.; Zhou, G.; Zhang, Q.; Tao, J.; Zhang, Y.; Lin, J. Aboveground mangrove biomass estimation in Beibu Gulf using machine learning and UAV remote sensing. *Sci. Total Environ.* **2021**, *781*, 146816. [CrossRef]
41. Liang, Y.; Kou, W.; Lai, H.; Wang, J.; Wang, Q.; Xu, W.; Wang, H.; Lu, N. Improved estimation of aboveground biomass in rubber plantations by fusing spectral and textural information from UAV-based RGB imagery. *Ecol. Indic.* **2022**, *142*, 109286. [CrossRef]
42. David, R.M.; Rosser, N.J.; Donoghue, D.N. Improving above ground biomass estimates of Southern Africa dryland forests by combining Sentinel-1 SAR and Sentinel-2 multispectral imagery. *Remote Sens. Environ.* **2022**, *282*, 113232. [CrossRef]
43. Zhou, Z. Feature Selection and Sparse Learning. In *Machine Learning*; Tsinghua University Press: Beijing, China, 2016; pp. 247–252.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.