forests

MDPI

Article

Remote Sensing Estimation of Forest Aboveground Biomass Based on Lasso-SVR

Ping Wang [1], Sanqing Tan [1,*], Gui Zhang [2], Shuang Wang [3] and Xin Wu [3]

[1] School of Forestry, Central South University of Forestry and Technology, Changsha 410004, China
[2] Key Laboratory of Digital Dongting Lake of Hunan Province, Changsha 410004, China
[3] National Forest Fire Prevention Virtual Simulation Experimental Teaching Center, Changsha 410004, China
* Correspondence: t19920990@csuft.edu.cn

Abstract: With the Lutou Forest Farm as the research area, the Lasso algorithm was used for characteristic selection, and the optimal combination of variables was input into the support vector regression (SVR) model. The most suitable SVR model was selected to estimate the aboveground biomass of the forest through the comparison of the kernel function and optimal parameters, and the spatial distribution map of the aboveground biomass in the study area was drawn. The significance analysis of special variables showed good correlations between forest aboveground biomass and each vegetation index. There was a more significant correlation with some remote sensing bands, a less significant correlation with some texture features, and a strong correlation with DEM in the terrain features. When the parameters C is 2 and g is 0.01, the SVR model has the highest precision, which can illustrate 73% of the forest aboveground biomass, with the validation set R^2 being 0.62. The statistical analysis of the results shows that the total aboveground biomass of the Lutou Forest Farm is 4.82×10^5 t. The combination of Lasso with the SVR model can improve the estimation accuracy of forest aboveground biomass, and the model has a strong generalization ability.

Keywords: aboveground biomass; remote sensing estimation; Lasso algorithm; support vector regression model

Citation: Wang, P.; Tan, S.; Zhang, G.; Wang, S.; Wu, X. Remote Sensing Estimation of Forest Aboveground Biomass Based on Lasso-SVR. *Forests* **2022**, *13*, 1597. https://doi.org/10.3390/f13101597

Academic Editors: Qisheng He and Wenmei Li

Received: 29 July 2022
Accepted: 23 September 2022
Published: 29 September 2022

1. Introduction

Forest aboveground biomass (AGB) is one of the critical parameters to assess forest ecosystem productivity and health status. It is of great significance to the global carbon cycle and climate change. With the progress and development of science and technology, a method for estimating forest aboveground biomass by remote sensing needs to be proposed to replace the traditional forest aboveground biomass estimation. The traditional forest aboveground biomass estimation is mainly completed by biomass survey methods, such as the harvesting method, standard wood method, model method [1], biomass conversion factor continuous method [2,3], etc., which consume a lot of manpower, material resources, and high research costs [4]. Additionally, in each study, the research area is limited, and thus, the efficiency is not high. In addition, as the data obtained are temporary, it is impossible to evaluate forest aboveground biomass whenever and wherever we want. Worse still, the spatial change of forest aboveground biomass is disturbed, which is not conducive to the growth of plants and also destroys the ecological environment. The balanced development of forest ecology is affected correspondingly.

In recent years, machine learning methods have become more and more widely used in the field of forestry. Among them, random forest (RF), support vector regression (SVR), BP neural network, multiple linear regression (MLR), deep learning, and other methods used in forest aboveground biomass estimations have been mentioned in some reports [5–7]. In both domestic and foreign studies, these methods have been adopted to classify, predict, and simulate forest vegetation types; vegetation coverage; and vegetation transition. For example, Zhang [8] used Landsat [5] TM images as research data. Additionally, multiple

linear regression (MLR) and BP neural network models were used to estimate the biomass of the Tahe and Amur Forest areas on the northern slope of Yilehuli Mountain in the northern part of Heilongjiang Province in China. Gleason et al. [9] compared the effects of linear mixed effects (LME) regression, random forest (RF), support vector regression (SVR), and cubism on estimating biomass in medium dense forests (with the canopy density between 40% and 60%) at the tree and sample levels. According to the study, it was found that when biomass estimation precision is enhanced, SVR produces the most accurate biomass models during modeling at the sample level. Zhang et al. [10] developed a new method for collaborative biomass estimation by integrating LiDAR data with Landsat8 images through a deep learning-based workflow. Using Landsat8 OLI as the data source, López-Serrano et al. [11] used the RF model and the SVR model to estimate the aboveground forest biomass observed in Madrid, Mexico, and the results showed that the SVR model is the best for estimating the aboveground forest biomass. Halme et al. [12] used two machine learning algorithms, Gaussian Process Regression (GPR) and SVR, to estimate forest biomass and structural variables in the northern coniferous forest of Finland. The study showed that the performance of GPR is slightly better than that of SVR in terms of estimating variables of basal areas and basal surface areas.

Since there are many vegetation indices, texture information, single-band information, and topographic factors that affect the estimation of aboveground forest biomass, the more characteristic variables that can be used for modeling, the greater the computational complexity of the estimation model. Additionally, when the number of selected characteristic variables is larger than the optimal number, the estimation accuracy may decrease. Therefore, there are some problems waiting to be solved: first, how do we select the optimal characteristic variable among numerous characteristics? Second, how do we establish a better machine learning model? Third, how do we estimate forest aboveground biomass effectively? Characteristic screening in machine learning can solve this problem efficiently. One study has shown that characteristic variable selection is very important to shorten the running time of the model and improve its accuracy. It is necessary to select a small and optimal characteristic set for modeling [13,14]. Wang [15] proposed that the Lasso algorithm can be used for characteristic selection and then used in the SVM model, which indicates that the use of the Lasso algorithm for characteristic variable selection is effective. The Lasso algorithm is a variable selection method based on coefficient compression. The general linear least squares method adds a constraint that requires the sum of the absolute values of all coefficients to be less than a certain constant, because this constraint may make some regression coefficients obtained through the regression model be zero, which facilitates the selection of variables and description of the model [16].

This study takes the Lutou Forest Farm as the research area, adopts forest resource planning and design survey data as the basic data, and uses geographic remote sensing information technology to extract a single band, vegetation index, texture information, and topographic characteristics. With the forest aboveground biomass as the estimation target, this study intends to establish an aboveground biomass estimation model of the Lutou Forest Farm and draw the spatial distribution mapping with the combination of the Lasso algorithm characteristic variable screening algorithm with SVR. Meanwhile, the degree of correlation between factors such as single band, vegetation index, texture characteristics, terrain, and the estimation of forest aboveground biomass is discussed in the study, providing a reference for the combination of machine learning methods and the selection of aboveground forest biomass variables.

2. Materials and Methods

2.1. Overview of the Research Area

Lutou Forestry Farm is an experimental forestry farm of Central South University of Forestry Technology. Lutou Forest Farm is located in Jiayi, Pingjiang County, Yueyang City, Hunan Province, China, between 113°51′52″~113°58′24″ east and 28°31′27″~28°38′00″ north latitude, covering a total of 4762 hectares, as shown in Figure 1 [17]. The terrain of

Lutou Forest Farm is high in the south and low in the north, with the main peak Shibazhe showing a north–south trend. The southern mountainous area is relatively low and flat, with high mountains and deep valleys in the middle and gentle mountains in the northern part. The highest area of the research area is the main peak Shibazhe, 1272.5 m above sea level, and the lowest area is on the back of the mountain, 124 m above sea level. The relative height difference is 1148.6 m. The landform types are mainly medium mountains and low mountains, but the mountains are steep, and the soil layer at the foot of the mountains is very thin. The forest farm belongs to the type of evergreen broad-leaved forest that is characterized by rich and diverse forest vegetation, mainly a secondary broad-leaved forest, Chinese fir plantations, and bamboo forest. There is a noticeable transition trend from East China to Central China and South China. It is a channel for plants to grow from south to north and from east to west. Inside the forest farm, there is the Castanopsis Eyrei Community with the largest distribution area, the lowest altitude, and the most complete structure so far.

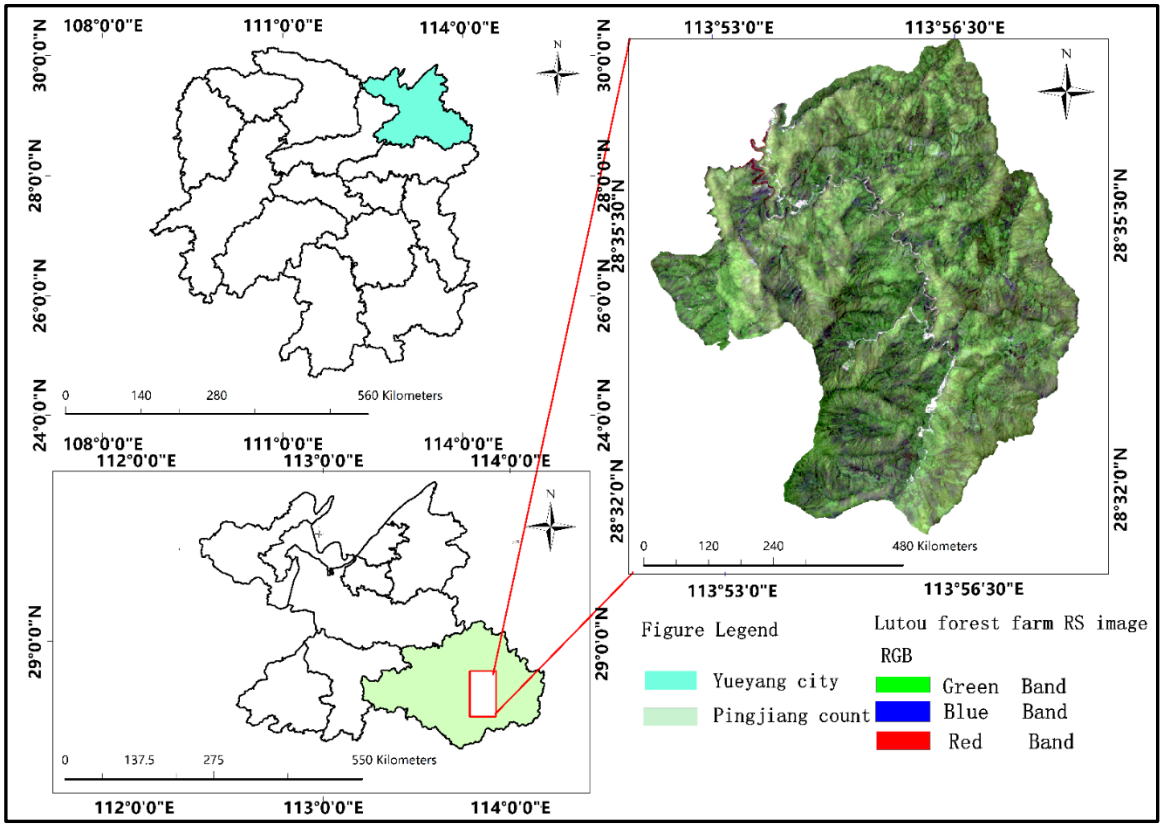

Figure 1. Location of the study area.

2.2. Ground Survey Data

The ground data of this study is derived from the survey data of forest resources planning and design in 2020. There are 1236 patches of land in total. In this study, one sample plot of 20 m × 20 m in size was taken from each small group. The data included tree height, diameter at breast height, tree species, altitude, grade, aspect, agrotype, soil depth, canopy density, forest categories, etc. See Table 1 [17]. The height and diameter at breast height (DBH) of trees with a DBH greater than 5 cm were obtained through tally, and then, according to aboveground biomass equations, the aboveground biomass of different tree species was calculated according to the tree species group in Table 2. Finally, the aboveground biomass of different tree species was added up as the aboveground biomass of the plot, and the sum was divided by the acreage of the sample plot. In the end, the aboveground biomass of the forest per unit area was calculated.

Table 1. Summary table of the fieldwork data.

ORIG_ FID	Area	Area (Acre)	Tree Species	Soil Type	Diameter at Breast Height	Tree Height	Number of Plants	Slope	Soil Thickness
0	1.64	25	310,000	103	6.00	10.00	1223	4	40
1	1.85	28	310,000	104	8.00	10.00	1950	3	60
2	4.27	64	310,000	103	8.00	10.00	2600	4	50
3	0.52	8	310,000	103	0.00	10.00	2100	4	40
4	18.18	273	310,000	103	14.00	10.00	2311	4	50
5	0.36	5	590,000	103	6.00	10.00	1415	4	40
6	1.01	15	310,000	103	14.00	10.00	1194	4	50
7	1.55	23	310,000	103	8.00	10.00	1061	4	50
8	2.81	42	310,000	103	14.00	10.00	1194	4	50
9	4.92	74	310,000	103	16.00	10.00	1238	4	50
10	0.27	4	310,000	103	18.00	10.00	1452	5	40
11	0.15	2	310,000	103	12.00	10.00	2746	4	50
12	0.48	7	590,000	103	10.00	10.00	1415	5	40
13									
14									
1236	0.48	7	590,000	103	18.00	10.00	1452	5	40

Table 2. Aboveground biomass calculation formula of different tree species.

Serial Number	Tree Species (Group)	Calculation Formula		References
1	Horsetail Pine	$W_S = 0.0237(D^2H)^{1.0015}$	$W_B = 0.0016(D^2H)^{1.1628}$	[18]
		$W_L = 0.0017(D^2H)^{1.0033}$;	$W_T = W_S + W_B + W_L$	[18]
2	Camphor Tree	$W_S = 0.0296(D^2H)^{0.9559}$;	$W_B = 0.0204(D^2H)^{0.8276}$	[18]
		$W_L = 0.0078(D^2H)^{0.8071}$;	$W_T = W_S + W_B + W_L$	[18]
3	Cedarwood	$W_S = 0.0422(D^2H)^{0.8623}$;	$W_B = 0.0206(D^2H)^{0.7367}$;	[18]
		$W_L = 0.0664(D^2H)^{0.5589}$;	$W_T = W_S + W_B + W_L$;	[18]
4	Oak	$W_S = 0.0560(D^2H)^{0.9140}$;	$W_B = 0.0080(D^2H)^{1.0370}$;	[18]
		$W_L = 0.0060(D^2H)^{0.8830}$;	$W_T = W_S + W_B + W_L$;	[18]
5	hard broad-leaved forest	$W_S = 0.0545(D^2H)^{0.8630}$;	$W_B = 0.0155(D^2H)^{0.8737}$;	[18]
		$W_L = 0.0145(D^2H)^{0.7444}$;	$W_T = W_S + W_B + W_L$;	[18]
6	soft broad-leaved forest	$W_S = 0.0699(D^2H)^{0.8254}$;	$W_B = 0.0267(D^2H)^{0.7207}$;	[18]
		$W_L = 0.0125(D^2H)^{0.6181}$;	$W_T = W_S + W_B + W_L$;	[18]
7	Bamboo	$W_T = 0.6439D^{1.5373}$;		[19]

Description: W_S: stem biomass, W_B: branch biomass, W_L: leaf biomass, W_T: aboveground biomass, D: diameter at breast height (DBH), and H: tree height.

2.3. Remote Sensing Data

The Sentinel2 remote sensing image data adopted in this study was downloaded from the standard product data of the L1C level of Sentinel2 image in April 2020 in the European Space Agency (ESA) Copernicus Data Center [20,21]. The Sentinel2 satellite image contains 13 bands of multispectral data. Among them, the spatial resolution of Band2, Band3, Band4, and Band8 is 10 m, which are the red band, green band, blue band, and near-infrared band, respectively [20]. The spatial resolution of Band5, Band6, Band7, Band8a, Band11, and Band12 is all 20 m, and the spatial resolution of the other three bands is 60 m [22].

By adopting the Sen2 cor plug-in in the SNAP software provided by ESA, the LIC-level standard product data of Sentinel2 was processed for atmospheric correction [23]. In each band of the corrected image, except for the four bands of blue, green, red, and near-infrared, the rest of the bands are sampled to their original resolutions; the four bands of blue, green, red, and near-infrared, except for the maintenance of their original 10-m resolution, were used for texture calculation and sampled to a 20-m resolution, consistent with the other six bands with a 20-m resolution. They were mainly used for band combination, index calculation, etc. [21]. Finally, through Layer Stacking in ENVI software, each band was combined into two groups of images. The bands with 10-m spatial resolution are Band2, Band3, Band4, and Band8; the ones with 20-m spatial resolution are Band2, Band3, Band4, Band5, Band6, Band7, Band8, Band8a, Band11, and Band12. The image data that has been geometrically corrected were selected as the reference image, and the more obvious and subtle intersections on the artificial surface were selected as control points. In addition, another raster file was registered, so that the same target object was presented in the same area of the corrected image. As a result, the precise geometric correction of the image of Lutou Forest Farm was realized. Finally, the Sentinel2 image was cropped based on the cropping tool according to the region of interest in ENVI5.3 (GeoScene Information Technology Co, Ltd., Beijing, China). In the end, the remote sensing image of Lutou Forest Farm was obtained [21].

2.4. Candidate Variables for Modeling

Forest AGB consists of four parts: stems, branches, and leaves, and there is a certain degree of correlation with the vegetation index, texture characteristics, single band, and geological landforms. In this study, through the combination of bands, texture information extraction, principal component analysis, and calculation of various vegetation indices of Sentinel2 remote sensing images (see the list of vegetation indices in Table 3 for the calculation formula), three groups (37 in total) of characteristic variables were extracted, including single band, vegetation index, and texture information. In terms of the texture information, the principal component operation of Band2, Band3, Band4, and Band8 in the Sentinel2 multispectral image produced a total of 11 new texture characteristic variables, which were P_{1MEAN}, P_{1VARI}, P_{2VARI}, P_{1CONT}, P_{2CORR}, P_{1CORR}, P_{1HOMO}, P_{1DISS}, P_{2DISS}, P_{1ENTR}, and P_{1SECD}.

Table 3. Appropriate vegetation indices for Sentinel image 2.

Serial Number	Vegetation Index	Acronym	Formula	References
1	Ratio vegetation index	RVI	NIR/R	[23]
2	Red edge ratio vegetation index	RVI_{re5}	$NIR/RE1$	[23]
3	Normalized difference vegetation index	$NDVI$	$(NIR-R)/(NIR+R)$	[24]
4	Normalized Difference Red Edge Band 5 Vegetation Index	$NDVI_{re5}$	$(NIR-RE1)/(NIR+RE1)$	[25]
5	Normalized Difference Red Edge Band 6 Vegetation Index	$NDVI_{re6}$	$(NIR-RE2)/(NIR+RE2)$	[26]
6	Modified Normalized Difference Vegetation Index	$mNDVI$	$(NIR-R)/(NIR+R-2B)$	[26]
7	Modified Normalized Difference Red Edge Vegetation Index	$mNDVI_{re5}$	$(NIR-RE1)/(NIR+RE1-2B)$	[26]
8	Normalized Difference Infrared Index	$NDII$	$(NIR-S)/(NIR+S)$	[27]

Table 3. *Cont.*

Serial Number	Vegetation Index	Acronym	Formula	References
9	Green Light Chlorophyll Vegetation Index	Cl_{green}	$NIR/G - 1$	[28]
10	Red Edge Chlorophyll Vegetation Index	Cl_{re5}	$(NIR/RE1) - 1$	[28]
11	Enhanced vegetation index	EVI	$2.5(NIR - R)/(NIR + 6R - 7.5B + 1)$	[29]
12	Modified simple ratio index	MSR	$((NIR/R) - 1)/\sqrt{(NIR/R) + 1}$	[30]
13	Difference vegetation index	DVI	$NIR - R$	[31]
14	Nonlinear index	NLI	$((NIR)^2 - R)/((NIR)^2 + R)$	[32]
15	Red edge nonlinear index	NLI_{re5}	$((NIR)^2 - RE1)/((NIR)^2 + RE1)$	[32]
16	Novel inverted red-edge chlorophyll index	$IRECI$	$(RE3 - R)/(RE1 - RE2)$	[33]

Description: Near-infrared (*NIR*), RED (*R*), Red Edge 1 (*RE1*), Red Edge 2 (*RE2*), Red Edge 3 (*RE3*), Blue (*B*), SWIR-1 (*S*), and GREEN (*G*).

2.5. NASA DEM Data

NASA DEM data are the new global 30-m resolution DEM data released by NASA on 18 February 2020. NASA DEM is the highest resolution, best quality, and widest coverage DEM product for the foreseeable future. The scene number selected for this study is NASADEM_HGT_n28e113, and the slope data, slope direction data, and elevation data are calculated from the DEM data, and the slope data, slope direction data, and elevation data are interpolated to maintain the same spatial resolution as the Sentinel2 data.

2.6. Method Flow Chart

The overall structure of the paper is to extract numerous remote sensing variables related to forest aboveground biomass by ENVI software. After a correlation analysis combined with LASSO for feature selection, a support vector regression forest aboveground biomass estimation model was constructed to produce a forest aboveground biomass distribution map of the study area, as shown in Figure 2.

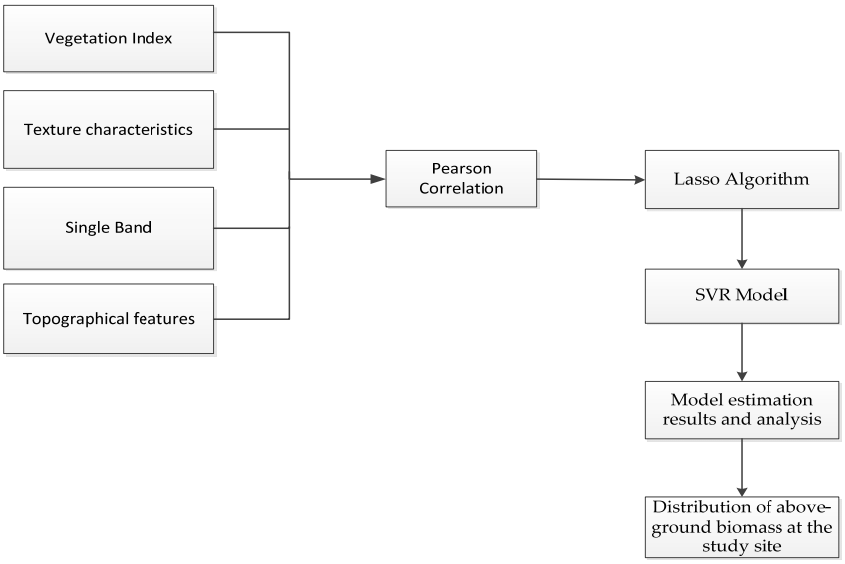

Figure 2. Method flow chart.

2.7. Pearson Correlation

A Pearson correlation coefficient is used to reflect the degree of linear correlation between two random variables so as to measure the linear correlation. The value of r is between -1 and 1. When the value is 1, it means that there is a completely positive correlation between the two random variables; when the value is -1, it means that there is a completely negative correlation between the two random variables; when the value is 0, it means that the two random variables are linearly independent [34]. The formula for calculating the Pearson correlation coefficient is as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (1)$$

Description: n refers to the sample quantity; X_i, and Y_i are the observed values of the i point corresponding to the variables X and Y; $\overline{X}$ is the mean value of the sample X; and $\overline{Y}$ is the mean value of the sample Y.

2.8. Lasso Algorithm

The Lasso algorithm is a variable selection method based on the coefficient compression method. Lasso regression adds constraints on the basis of the general linear least squares method, which requires a certain constant to be greater than the sum of the absolute values of the coefficients. The constraints are likely to make some regression coefficients obtained through the regression model zero, which is beneficial to the selection of variables and description of the model. Lasso regression models not only have the advantage of being easy to be interpreted as the optimal subset selection but also have a stability advantage similar to ridge regression [35].

The basic idea of Lasso regression is to get the estimated value of the regression coefficient. If a certain value is greater than the absolute sum of the regression coefficient, the residual sum of squares of the regression equation will be minimized. Lasso regression can effectively and quickly reduce the data dimensions, which is extremely suitable for the variable selection of high-dimensional data [35]. The mathematical description of LASSO regression equivalence is as follows:

$$argmin\left\{\sum_{i=1}^{n}\left(y_i - u_0 - \sum_{j=1}^{P} u_j x_{ij}\right)^2 + \lambda\sum_{j=0}^{p}|\mu_j|\right\} \quad (2)$$

Description: x_{ij} represents the independent variable of the sample sequence i, and y_i represents the dependent variable of the sample sequence i. $\lambda\sum_{j=0}^{p}|u_j|$ is the penalty function. λ represents the penalty parameter of the model. The larger the value of λ, the more variables are deleted; on the contrary, the smaller the value of λ is, the less variables are deleted [35]. In this paper, the R language program was used to control the coefficient lambda value before the penalty term of the Lasso model to obtain the optimal model and the variable coefficients and optimal variables of the model.

Since Lasso is equipped with the function of variable screening, it can be used when selecting the aboveground biomass variables of the forest. As a result, a high-precision remote sensing estimation model of forest aboveground biomass based on Lasso-SVR can be formed. The Lasso algorithm not only screened the variables most related to aboveground forest biomass but also optimized the support vector regression model and improved the estimation accuracy of aboveground biomass in the research area.

2.9. SVR Model

The SVR model is used to solve the binary classification problem, and it is widely used to support vector machines in the field of regression. What the SVR model seeks is a linear regression equation model that is suitable for all sample points, and the optimal

hyperplane it seeks is the total variance of the minimized sample points from the hyperplane instead of dividing the two classes [36]. The SVR model constructs an error range, treats the predicted value within the error range as a correct prediction, and establishes a regression model according to the size of the given error interval. The training samples $\{(x_i, y_i), i = 1, 2, \ldots, n\}$ were used as the input variables of the model. x was the predicted value, and y was the measured value. The standard form of the SVR model is:

$$\min_{\omega,b} \frac{1}{2}\|\omega^2\| + C\sum_{i=1}^{m} L_\varepsilon(y - f(x)) \quad (3)$$

$$L_\varepsilon(y - f(x)) = \begin{cases} 0 & |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon & \text{other} \end{cases} \quad (4)$$

In the formula: $f(x)$ represents the output model; ω represents the normal vector, ε represents the insensitive loss, L_ε represents the insensitive loss function, and C represents the model penalty coefficient. As the accuracy of model training has a positive correlation with the C value, the overfitting of the model may occur when the C value is far beyond the normal range.

SVR also includes kernel tricks, with an aim to make data separable in high-dimensional space by mapping linearly inseparable data in the input space to the high-dimensional characteristic space. Therefore, the sample inner product needs to be calculated in this process. Since there is an excessively large sample dimension, a kernel function needs to be introduced to convert the high-dimensional vector inner product calculation into a low-dimensional vector inner product calculation. Sigmoid function, radial basis function, polynomial function, and linear function all belong to commonly used kernel functions of SVR.

The accuracy evaluation of the model requires certain model evaluation indicators. The model indicators directly reflect the degree of fitting of the model and its quality. In this study, 3 indicators, including the coefficient of determination R^2, root mean square error (RMSE), and mean absolute error (MAE), evaluate the model accuracy [35].

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (5)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (7)$$

Among them: the measured value of the sample biomass (the sequence i) is y_i, the mean of the measured results is $\overline{y}$, the estimated value of the sample biomass (the sequence i) is $\hat{y}_i$, and the number of samples is n.

This study used the glmnet package and the e1071 package in the R language environment to estimate the aboveground biomass on the Lutou Forest Farm. (1) Sample data (1236) were selected as the experimental data, of which the number of samples in the modeling set was 865, and the number of samples in the validation set was the remaining 371. (2) The glmnet software package was utilized to screen the 28 initial characteristic after the Pearson correlation analysis, and the selected optimal variable was selected as the estimated variable. (3) The e1071 software package was employed to build the SVR model and optimize the parameters, and finally, the estimation results of the biomass on the Lutou Forest Farm were evaluated.

3. Results and Analysis

3.1. Correlation Analysis

Generally, each type of forest AGB has different forest structures and characteristics. Therefore, 16 vegetation indices, 11 texture features, 10 single-band features, and 3 to-

pographic variables, a total of 40 feature variables, were selected as the remote sensing preparatory variables for estimating forest AGB. SPSS25.0 software (Chicago, IL, USA) was used to conduct the correlation degree analysis study of forest AGB in the form of feature variable groups. The results of the correlation are shown in Figure 3. The results indicate that the correlation between forest AGB and each vegetation index is good, and there is a more significant correlation between some remote sensing bands, a less significant correlation with some texture features, and a strong correlation with the topographic features. From the specific cases, the topographic features DEM; single-bands Band2, Band3, Band4, Band5, Band6, Band11, and Band12; texture features P_{1ENTR}, P_{1SECD}, P_{1VARI} , P_{2VARI}, P_{1CONT}, P_{1CORR}, P_{2CORR}, and P_{1DISS}; and vegetation indices *NDVI*, $NDVI_{re5}$, $NDVI_{re6}$, *RVI*, RVI_{re5}, Cl_{green}, *EVI*, *mNDVI*, *MSR*, $mNDVI_{re5}$, NLI_{re5}, and *IRECI* were all highly correlated with forest AGB, and the relationships were significant. Therefore, 12 vegetation indices, 8 texture features, 7 single-band variables, and 1 topographic variable were selected to have high correlations with forest AGB, totaling 28 variables.

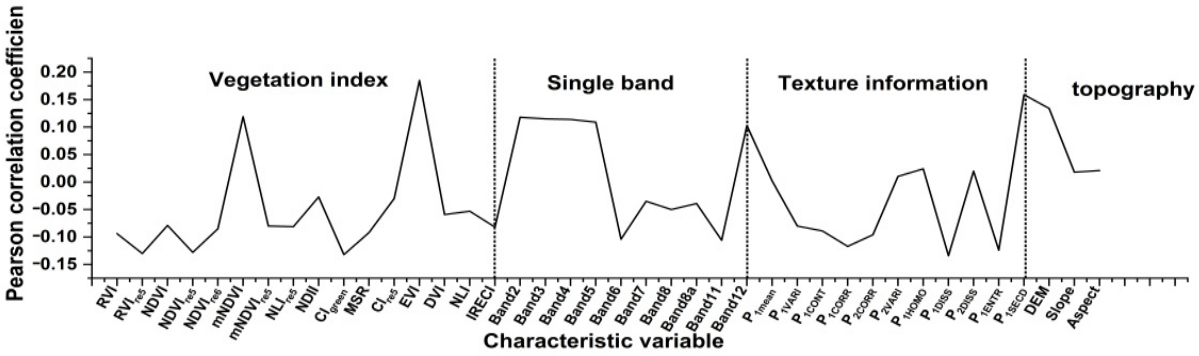

Figure 3. Pearson's correlation between characteristic variables and forest aboveground biomass.

3.2. Lasso Algorithm Feature Selection

Aiming at the different dimensions of the data after the correlation analysis, this study chose the normalization method to process the data. The data were converted to a decimal between 0 and 1. The method of solving Lasso regression is the glmnet package of R language. The glmnet package controls the variables through the parameter lambda value and selects the model with the best and least number of independent variables. The larger the lambda value, the more variables are eliminated; on the contrary, the smaller the lambda value, the fewer variables are eliminated [33].

According to Figure 4, it can be seen that, as the lambda value gradually increases, the variable coefficients are continuously compressed to 0, and the degrees of freedom and residuals also gradually become smaller. In Figure 4, it is obvious that the *MSR*, P_{1CORR}, *NDVI*, P_{1ENTR}, Band12, and $NDVI_{re6}$ variables gradually become 0 as the lambda value increases, indicating that they play a key role in estimating the aboveground biomass of the forest (Figure 4). The red dots in Figure 5 indicate the target parameters corresponding to each lambda value, and the two dashed lines indicate the special lambda values. The left dashed line refers to the minimum mean square error value within all lambda values, and the right dashed line refers to the lambda value for the simplest model where the mean square error reaches a minimum level within a range of variance. That is, when the lambda value is equal to 0.0047, then the minimum model can be obtained. A total of 13 variables are selected for this model, including *NDVI*, $NDVI_{re6}$, Cl_{green}, *mNDVI*, *MSR*, *IRECI*, DEM, Band4, Band5, Band12, P_{1ENTR}, P_{1CORR}, and P_{1DISS} as independent variables. The results of the Lasso feature selection are shown in Figure 6.

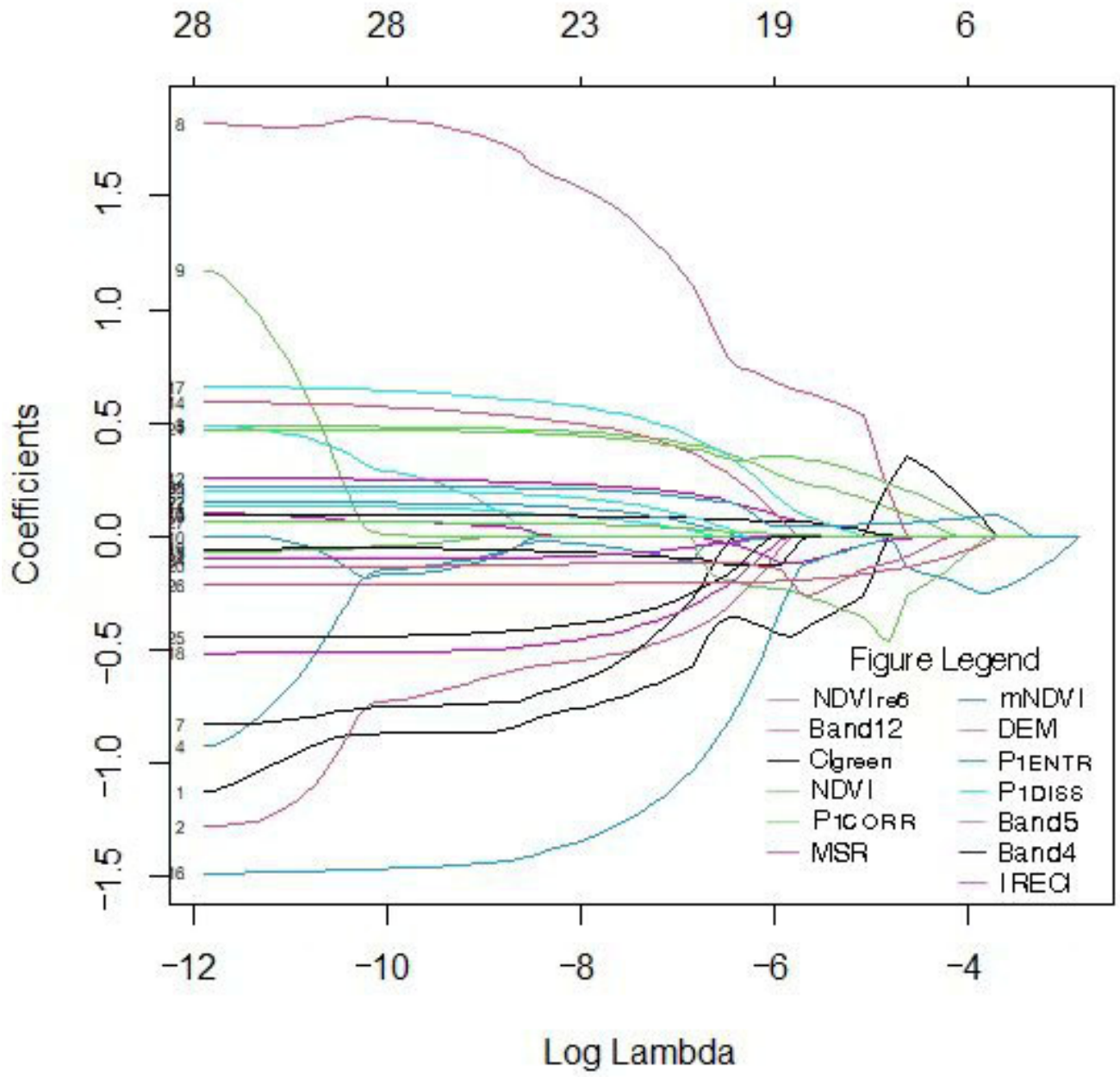

Figure 4. Glmnet package filter variables.

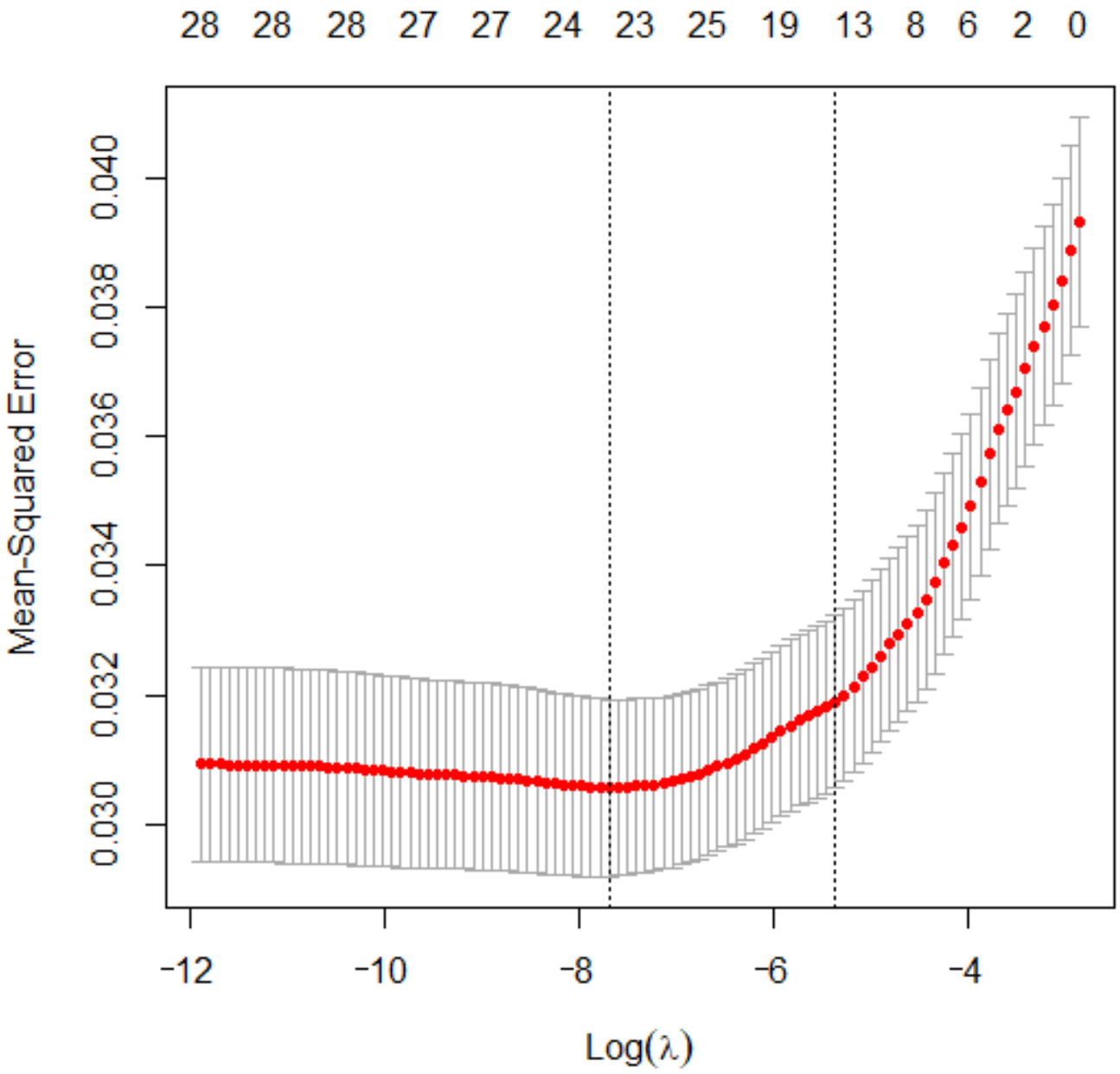

Figure 5. Glmnet package filter variable process Lambda values.

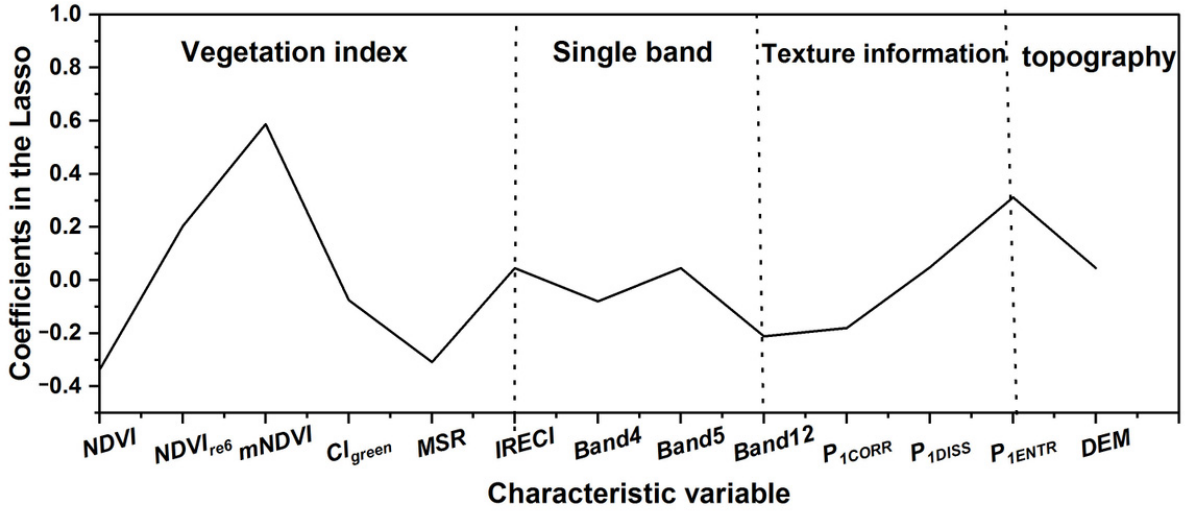

Figure 6. Lasso feature selection results.

According to the figure of feature selection result graph, the following conclusions can be reached: (1) The aboveground biomass estimation in the research is most closely correlated with the vegetation index. The normalized vegetation index series is the primary correlation factor, followed by the chlorophyll series index and, finally, the modified simple vegetation index. This is due to the fact that the Lutou Forest Farm has been operated for a long time and belongs to evergreen broad-leaved forest. The vegetation types are rich and diverse, and the plants there are lush and tall. As $NDVI$, $NDVI_{re6}$, CI_{green}, $mNDVI$, MSR, and $IRECI$ are sensitive to the green vegetation index, they are used to describe the vegetation profile and development status. Therefore, they are most closely correlated to forest aboveground biomass. (2) The altitude disparity of Lutou Forest Farm varies greatly. Usually, the area with high terrain is rarely visited by people. Affected by the changing course of the four seasons, the aboveground biomass of branches, leaves, and stems is scarce. Therefore, it can be concluded that the altitude is relatively correlated with the forest aboveground biomass. (3) According to the spectral characteristics of plants, different organisms will generate different reflectivity and absorptivity for different wavelengths, and the final reflection on the image will be different. Band4 and Band5 are used to monitor the growth of plants, and Band12 is used to distinguish live biomass, dead biomass, and soil. As a result, Band4, Band5, and Band12 also have a certain degree of correlation with forest aboveground biomass. (4) P_{1ENTR}, P_{1CORR}, and P_{1DISS} are important parameters. They are mainly used to identify the attributes of the ground objects due to its convenience in distinguishing the biomass. Therefore, the texture information is related to the forest aboveground biomass, but the correlation is not high.

In order to verify whether the precision of the model has been improved after the Lasso algorithm screening, this study used ten-fold cross-validation to conduct two experiments with the SVR model of the initial characteristic variables and the variables after the Lasso algorithm feature selection. With the modeling set and the validation set MAE, RMSE, and R^2 as the schedule indices, the characteristic variables in the SVR model with the highest precision are selected as the final characteristic variables of modeling. As shown in Table 4, the modeling set R^2 using the initial characteristic variables is 0.75, the validation set R^2 is 0.60, the modeling set R^2 using the Lasso characteristic variables after is 0.73, and the validation set R^2 is 0.62. In the validation set, the R^2 using the initial feature variables is lower than the R^2 using the significant feature variables after Lasso screening, where the RMSE and MAE are also lower than the optimal feature variables. The modeling set R^2 and the validation set R^2 of the feature variables after selection using the Lasso algorithm are close and the model is relatively stable. The results show that it is effective to adopt characteristic screening and use it in SVR modeling. Therefore, a total of 13 variables were used, including the $NDVI$, $NDVI_{re6}$, CI_{green},

mNDVI, *MSR*, *IRECI*, DEM, Band4, Band5, Band12, P_{1ENTR}, P_{1CORR}, and P_{1DISS} to estimate the aboveground biomass of the research area.

Table 4. Final verification results of feature variable screening.

Characteristic Variable	Modeling Set			Validation Set		
	RMSE (t/ha)	MAE (t/ha)	R^2	RMSE (t/ha)	MAE (t/ha)	R^2
Original characteristic variable	29.58	19.85	0.75	36.46	24.79	0.60
Lasso characteristic variable	32.34	24.78	0.73	34.76	24.61	0.62

3.3. SVR Model Kernel Function and Parameter Selection

To determine the optimal feature selection variable, it is necessary to select the appropriate kernel function and its related parameters to ensure the modeling accuracy [36]. In this study, four commonly used kernel functions are adopted to establish the SVR model when the parameters are optimal. With the modeling set and the validation set R^2 as the standard to measure the accuracy of the model, the parameters of the SVR model are optimized while contrasting the modeling accuracy of different kernel functions. C and g are important parameters in the kernel function of the SVR model (except for the linear kernel function) [34]. Assigning the values of parameters C and g in logarithmic or exponential form tunes better than integers.

Call the tune.svm function in the e1071 software package to program, set the ranges of C and g within 10^{-2}~10^{0} and 10^{0}~10^{2}, respectively, and use the grid search method and ten-fold intersection to determine the optimal parameter [36]. The test results of different kernel functions of the SVR model in Table 5 show that the R^2 of selecting the Polynomial Kernel and the Radial Basis Function as the modeling set and validation set kernel function is relatively higher, and the validation set R^2 of selecting the Linear Kernel and Sigmoid Kernel is relatively lower [37,38]. In the forest aboveground biomass estimation model, there are many types of input characteristic variables, and the Radial Basis Function can better solve the linear inseparability problem of data aggregation. Therefore, this study selected the RBF function as the kernel function (the optimal parameters are C = 2 and g = 0.01) for modeling, and the results are similar to those of previous studies.

Table 5. Comparison of the test results of different kernel functions of the SVR model.

Kernel Function	Cost	Gamma	Number of Support Vector Machines	Modeling Set	Validation Set
				R^2	R^2
Radial Basis Function	2	0.01	766	0.73	0.62
Polynomial Kernel	1	3	776	0.63	0.55
Sigmoid Kernel	1	0.077	848	0.57	0.51
Linear Kernel	1	None	773	0.65	0.53

3.4. Estimation and Accuracy Evaluation of Forest Aboveground Biomass

The accuracy evaluation results (shown in Table 6 and Figure 7) show that the SVR model using the optimal combination of characteristic variables has a good degree of fitting between the estimated value of forest aboveground biomass and the measured value. The final SVR model can be used to explain 73% of the aboveground forest biomass, and the determination coefficient in the modeling set is similar to that in the validation set. The model

is relatively stable and has avoided the overfitting problem. The RMSE in the validation set is 34.76 t/ha, and the MAE is 24.61 t/ha. Though they are slightly higher than the modeling set, the accuracy is still high, indicating that the SVR model has high accuracy in the estimation of aboveground forest biomass and has good generalization ability.

Table 6. Forest aboveground biomass Lasso-SVR model accuracy validation.

Model	Modeling Set			Validation Set		
	RMSE (t/ha)	MAE (t/ha)	R^2	RMSE (t/ha)	MAE (t/ha)	R^2
Lasso + SVR	32.34	24.78	0.73	34.76	24.61	0.62

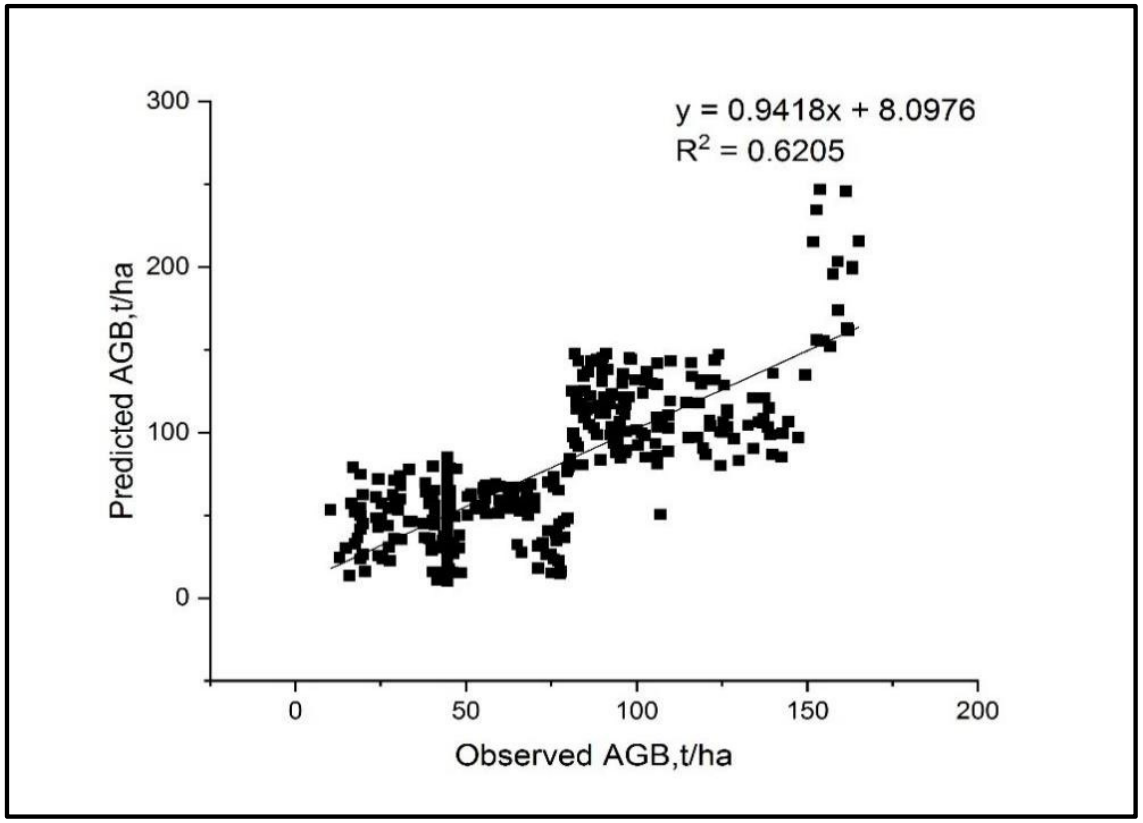

Figure 7. Validation results based on the Lasso-SVR model accuracy.

3.5. Mapping of Forest Aboveground Biomass Estimation

In this study, the optimal model obtained by the Lasso-SVR method was selected for the inversion of aboveground forest biomass in the study area, and the results are shown in Figure 8. From the statistical analysis of the regional inversion results, the total forest AGB in the study area was $4.82 \times 10^5 t$. The forest aboveground biomass was mainly distributed in the northwestern and central parts of the gently sloping forest land and gradually became more and more dense from south to north, which was also consistent with the field survey.

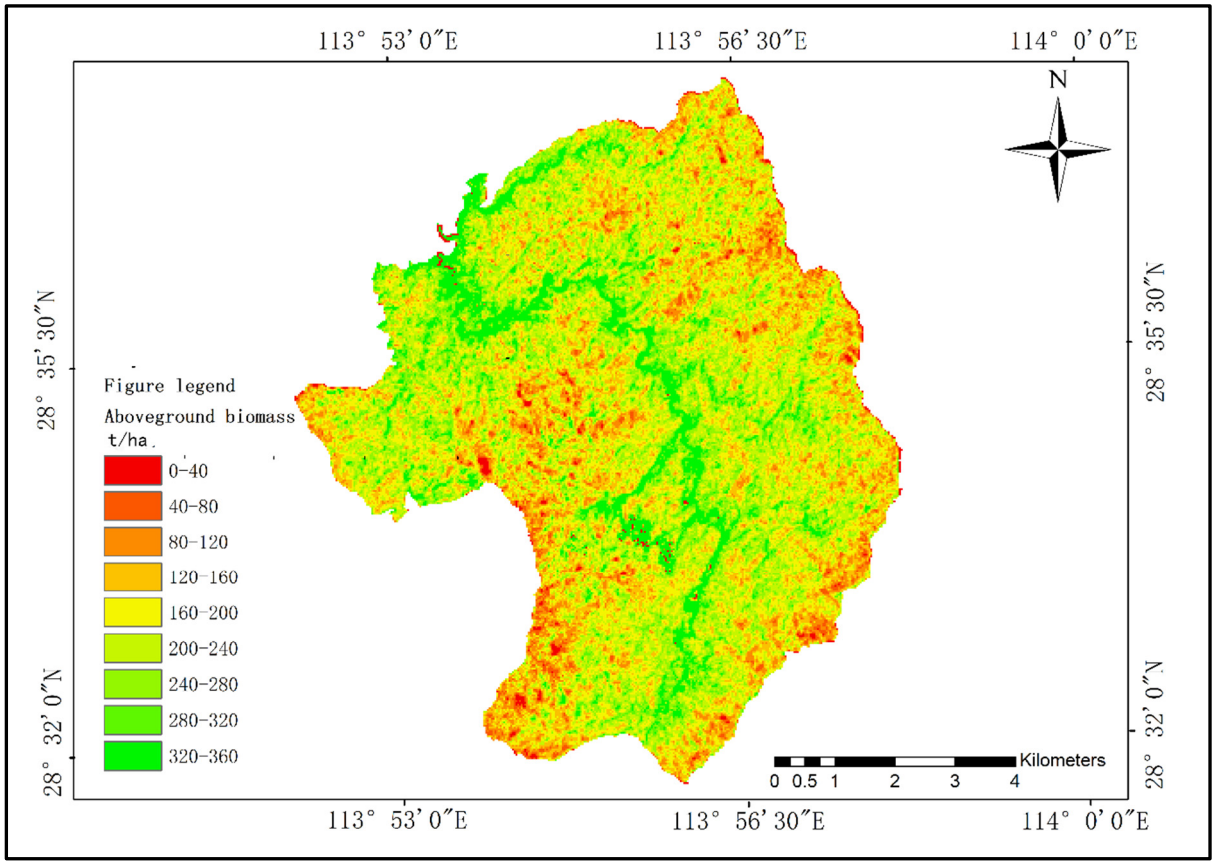

Figure 8. Forest aboveground biomass distribution of the Lutou Forest Farm.

4. Discussion

This paper study used remote sensing data and its derived characteristic factors (such as vegetation index, texture, single band, terrain factor, etc.), so that the estimation accuracy of forest aboveground biomass can be improved to a certain extent, but the use of remote sensing data and its derived characteristic factors is usually accompanied by problems of data in the higher dimensions, redundant information, and overfitting of the estimation models. Currently, there is still a lack of a stable and efficient feature selection method for forest aboveground biomass remote sensing modeling in this field. To effectively use high-dimensional remote sensing features for forest aboveground biomass estimations, this paper proposes an optimized SVR model based on Lasso feature selection. However, the following problems still exist. (1) Since the Lasso algorithm is not compared with other feature selection algorithms, the advantages of the Lasso algorithm in the feature selection algorithm cannot be better reflected. (2) According to the degree of correlation between the forest aboveground biomass and many other variables, when estimating the aboveground forest biomass, more relevant characteristic variables can be introduced to improve the final accuracy of the aboveground biomass estimation in the study area. (3) The Lasso-SVR model ignored the regionality of forest aboveground biomass. Additionally, the Lasso-SVR model is not used to estimate the aboveground biomass in more research areas, and a further comparative analysis was not made. Therefore, the advantages and disadvantages of the Lasso algorithm compared with other machine learning feature selection algorithms should be further contrasted and analyzed.

5. Conclusions

In this study, the optimal characteristic combination screened by the Lasso algorithm based on characteristic variable groups such as the vegetation index, texture factor, terrain factor, and single band was used as the estimated variable for the SVR model. In addi-

tion, different kernel functions and optimized parameters were compared, the Lasso-SVR estimation model for aboveground forest biomass was established, and a spatial distribution mapping was drawn. The following conclusions were obtained: (1) There are a total of 13 variables that are strongly correlated with aboveground biomass in the research area. They were the $NDVI$, $NDVI_{re6}$, Cl_{green}, $mNDVI$, MSR, $IRECI$, DEM, Band4, Band5, Band12, P_{1ENTR}, P_{1CORR}, and P_{1DISS}. (2) The combination of the Lasso algorithm with the SVR model for the mapping of aboveground forest biomass estimation can effectively enhance the accuracy of the model. Different combinations of characteristic variables have explained 73% of the aboveground forest biomass in the research area. The validation set R^2 is 0.62. (3) The SVR model kernel function and parameter selection show that, for the estimation of forest aboveground biomass, since most of the characteristic variables are linearly inseparable, it is most suitable to choose the RBF kernel function. For parameter selection, the SVR model has fewer adjustable and simple parameters.

Author Contributions: S.T. conceived and designed the study. P.W. wrote the first draft, performed the data analysis, and collected all the study data. S.W., G.Z. and X.W. provided critical insights in editing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Science and Technology Innovation Platform and Talent Plan Project of Hunan Province under Grant 2017TP1022, in part by the National Natural Science Foundation of China Youth Project 32201552, in part by the Philosophy and Social Science Foundation Youth Project of Hunan Province under Grant 21YBQ054.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

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

References

1. Saint-André, L.; M'Bou, A.T.; Mabiala, A.; Mouvondy, W.; Jourdan, C.; Roupsard, O.; Deleporte, P.; Hamel, O.; Nouvellon, Y. Age-related equations for above- and below-ground biomass of a Eucalyptus hybrid in Congo. *For. Ecol. Manag.* **2005**, *205*, 199–214. [CrossRef]
2. Zhang, X.Q.; Xu, D. Calculating forest biomass changes in China. *Science* **2002**, *296*, 1359. [CrossRef] [PubMed]
3. Fang, J.Y.; Wang, Z.M. Forest biomass estimation at regional and global levels, with special reference to China's forest biomass. *Ecol. Res.* **2001**, *16*, 587–592. [CrossRef]
4. Wijaya, A.; Kusnadi, S.; Gloaguen, R.; Heilmeier, H. Improved strategy for estimating stem volume and forest biomass using moderate resolution remote sensing data and GIS. *J. For. Res.* **2010**, *21*, 1–12. [CrossRef]
5. Santi, E.; Paloscia, S.; Pettinato, S.; Cuozzo, G.; Padovano, A.; Notarnicola, C.; Albinet, C. Machine-Learning Applications for the Retrieval of Forest Biomass from Airborne P-Band SAR Data. *Remote Sens.* **2020**, *12*, 804. [CrossRef]
6. Wu, C.; Shen, H.; Shen, A.; Deng, J.; Gan, M.; Zhu, J.; Xu, H.; Wang, K. Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery. *J. Appl. Remote Sens.* **2016**, *10*, 35010. [CrossRef]
7. Dang, A.T.N.; Nandy, S.; Srinet, R.; Luong, N.V.; Ghosh, S.; Kumar, A.S. Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. *Ecol. Inform.* **2019**, *50*, 24–32. [CrossRef]
8. Zhang, Y.Y.; Li, F.R.; Liu, F.X. Forest biomass estimation based on remote sensing method for north Daxingan moun-tains. In *Advanced Materials Research*; Trans Tech Publications Ltd.: Zurich, Switzerland, 2011; Volume 339, pp. 336–341. [CrossRef]
9. Gleason, C.J.; Im, J. Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sens. Environ.* **2012**, *125*, 80–91. [CrossRef]
10. Zhang, L.; Shao, Z.; Liu, J.; Cheng, Q. Deep Learning Based Retrieval of Forest Aboveground Biomass from Combined LiDAR and Landsat 8 Data. *Remote Sens.* **2019**, *11*, 1459. [CrossRef]
11. López-Serrano, P.M.; Cárdenas Domínguez, J.L.; Corral-Rivas, J.J.; Jiménez, E.; López-Sánchez, C.A.; Vega-Nieva, D.J. Modeling of aboveground biomass with Landsat 8 OLI and machine learning in temperate forests. *Forests* **2019**, *11*, 11. [CrossRef]
12. Halme, E.; Pellikka, P.; Mõttus, M. Utility of hyperspectral compared to multispectral remote sensing data in estimating forest biomass and structure variables in Finnish boreal forest. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *83*, 101942. [CrossRef]
13. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
14. Nilsson, R.; Pena, J.M.; Björkegren, J.; Tegnér, J. Consistent feature selection for pattern recognition in polynomial time. *J. Mach. Learn. Res.* **2007**, *8*, 589–612.

15. Wang, K.; Liu, L.; Yuan, C.; Wang, Z. Software defect prediction model based on LASSO–SVM. *Neural Comput. Appl.* **2021**, *33*, 8249–8259. [CrossRef]
16. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* **1996**, *58*, 267–288. [CrossRef]
17. He, H.; Zhu, G.; Ma, W.; Liu, F.; Zhang, X. Additivity of stand basal area predictions in canopy stratifications for natural oak forests. *For. Ecol. Manag.* **2021**, *492*, 119246. [CrossRef]
18. Zhou, G.Y.; Yin, G.C.; Tang, X.L. *Carbon Stocks in China's Forest Ecosystems: A Biomass Equation*; Science Publishers: BeiJing, China, 2018; p. 70.
19. Li, H.K.; Lei, Y.C. *Assessment of Forest Vegetation Biomass and Carbon Stocks in China*; Chinese Forestry Press: Beijing, China, 2010; p. 58.
20. Phiri, D.; Simwanda, M.; Salekin, S.; Nyirenda, V.R.; Murayama, Y.; Ranagalage, M. Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sens.* **2020**, *12*, 2291. [CrossRef]
21. Chen, Y.; Guerschman, J.; Shendryk, Y.; Henry, D.; Harrison, M. Estimating Pasture Biomass Using Sentinel-2 Imagery and Machine Learning. *Remote Sens.* **2021**, *13*, 603. [CrossRef]
22. Raiyani, K.; Gonçalves, T.; Rato, L.; Salgueiro, P.; da Silva, J.M. Sentinel-2 Image Scene Classification: A Comparison between Sen2Cor and a Machine Learning Approach. *Remote Sens.* **2021**, *13*, 300. [CrossRef]
23. Muhsoni, F.F.; Sambah, A.B.; Mahmudi, M.; Wiadnya, D.G.R. Comparison of different vegetation indices for assessing mangrove density using sentinel-2 imagery. *GEOMATE J.* **2018**, *14*, 42–51. [CrossRef]
24. Yuan, F.; Bauer, M.E. Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. *Remote Sens. Environ.* **2007**, *106*, 375–386. [CrossRef]
25. Gitelson, A.A.; Merzlyak, M.N. Remote estimation of chlorophyll content in higher plant leaves. *Int. J. Remote Sens.* **1997**, *18*, 2691–2697. [CrossRef]
26. Sims, D.A.; Gamon, J.A. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sens. Environ.* **2002**, *81*, 337–354. [CrossRef]
27. Wilson, N.R.; Norman, L.M. Analysis of vegetation recovery surrounding a restored wetland using the normalized difference infrared index (NDII) and normalized difference vegetation index (NDVI). *Int. J. Remote Sens.* **2018**, *39*, 3243–3274. [CrossRef]
28. Gitelson, A.A.; Gritz, Y.; Merzlyak, M.N. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* **2003**, *160*, 271–282. [CrossRef]
29. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195–213. [CrossRef]
30. Chen, J.M. Evaluation of vegetation indices and a modified simple ratio for boreal applications. *Can. J. Remote Sens.* **1996**, *22*, 229–242. [CrossRef]
31. Naji, T.A.H. Study of vegetation cover distribution using DVI, PVI, WDVI indices with 2D-space plot. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2018; Volume 1003, p. 012083. [CrossRef]
32. Goel, N.S.; Qin, W. Influences of canopy architecture on relationships between various vegetation indices and LAI and FPAR: A computer simulation. *Remote Sens. Rev.* **1994**, *10*, 309–347. [CrossRef]
33. Frampton, W.J.; Dash, J.; Watmough, G.; Milton, E.J. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS J. Photogramm. Remote Sens.* **2013**, *82*, 83–92. [CrossRef]
34. Liu, Y.; Mu, Y.; Chen, K.; Li, Y.; Guo, J. Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient. *Neural Process. Lett.* **2020**, *51*, 1771–1787. [CrossRef]
35. Fonti, V.; Belitser, E. Feature selection using lasso. *VU Amst. Res. Pap. Bus. Anal.* **2017**, *30*, 1–25.
36. Marabel, M.; Alvarez-Taboada, F. Spectroscopic determination of aboveground biomass in grasslands using spectral transformations, support vector machine and partial least squares regression. *Sensors* **2013**, *13*, 10027–10051. [CrossRef] [PubMed]
37. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef] [PubMed]
38. Wang, H.; Xu, D. Parameter Selection Method for Support Vector Regression Based on Adaptive Fusion of the Mixed Kernel Function. *J. Control Sci. Eng.* **2017**, *2017*, 3614790. [CrossRef]