remote sensing

MDPI

Article

Aboveground Biomass Prediction of Arid Shrub-Dominated Community Based on Airborne LiDAR through Parametric and Nonparametric Methods

Dongbo Xie [1,2], Hongchao Huang [1,2], Linyan Feng [1,2], Ram P. Sharma [3], Qiao Chen [1], Qingwang Liu [1] and Liyong Fu [1,2,*]

[1] Research Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China
[2] Key Laboratory of Forest Management and Growth Modelling, National Forestry and Grassland Administration, Beijing 100091, China
[3] Institute of Forestry, Tribhuwan Univeristy, Kritipur 44600, Nepal
* Correspondence: fuly@ifrit.ac.cn; Tel.: +86-10-62889126

Abstract: Aboveground biomass (AGB) of shrub communities in the desert is a basic quantitative characteristic of the desert ecosystem and an important index to measure ecosystem productivity and monitor desertification. An accurate and efficient method of predicting the AGB of a shrub community is essential for studying the spatial patterns and ecological functions of the desert region. Even though there are several entries in the literature on the AGB prediction of desert shrub communities using remote sensing data, the applicability and accuracy of airborne LiDAR data and prediction methods have not been well studied. We first extracted the elevation, density and intensity variables based on the airborne LiDAR, and then sample plot-level AGB prediction models were constructed using the parametric regression (nonlinear regression) and nonparametric methods (Random Forest, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting Machine, and Multivariate adaptive regression splines). We evaluated accuracies of all the AGB prediction models we developed based on the fit statistics. Results showed that: (1) the elevation, density and intensity variables obtained from LiDAR point cloud data effectively predicted the AGB of the desert shrub community at a sample plot level, (2) the kappa coefficient of nonlinear mixed-effects (NLME) model obtained was 0.6977 with an improvement by 13% due to the random effects included into the model, and (3) the nonparametric model, such as Support Vector Machine showed the best fit statistics (R^2 = 0.8992), which is 28% higher than the NLME-model, and effectively reduced the heteroscedasticity. The AGB prediction model presented in this paper, which is based on the airborne LiDAR data and machine learning algorithm, will provide a valuable tool to the managers and researchers for evaluating desert ecosystem productivity and monitoring desertification.

Keywords: aboveground biomass; LiDAR; shrub community; desert; nonparametric methods

Citation: Xie, D.; Huang, H.; Feng, L.; Sharma, R.P.; Chen, Q.; Liu, Q.; Fu, L. Aboveground Biomass Prediction of Arid Shrub-Dominated Community Based on Airborne LiDAR through Parametric and Nonparametric Methods. *Remote Sens.* **2023**, *15*, 3344. https://doi.org/10.3390/rs15133344

Academic Editor: Henning Buddenbaum

Received: 14 May 2023
Revised: 24 June 2023
Accepted: 26 June 2023
Published: 30 June 2023

1. Introduction

Desertification is one of the most threatening regional environmental problems at present, which leads to the decline of vegetation productivity, severe soil degradation and the release of carbon stored in the soil and vegetation into the atmosphere [1]. As one of the most affected countries by desertification and severe wind and sand damage, China has a desert area of 261.16 million sq km, or 27.2% of the total land area of the country. The aboveground biomass (AGB) of the desert shrubs is the essential quantitative characteristic of a desert ecosystem and an important index to measure ecosystem productivity and desertification monitoring [2–4]. Although the vegetation biomass per unit area in the desert is relatively low, it plays a significant role in monitoring the trend of global carbon

sequestration, essential ecological services, and national green space carbon sequestration [3,5,6]. The root systems of desert shrubs penetrate deep into the soil, facilitating the accumulation of organic matter and nutrients, and with the increase of vegetation biomass, soil carbon sequestration capacity would increase [7,8]. Thus, an accurate estimate of the AGB of a desert plant community would be vital to analyzing the global trends of carbon sequestration.

Field sampling is the most reliable way to obtain the AGB of the desert plant community, but it is inefficient and time-consuming [9] as it requires destructive sampling and measurement, which could lead to the instability of a desert ecosystem [10]. Although destructive sampling could be avoided in subsequent monitoring by allometric AGB modeling based on the measured data, it has some limitations in vegetation monitoring on a large scale [10]. The development of remote sensing technology has filled the gap of the traditional methods in time and space, as this can achieve systematic, efficient, and continuous monitoring at different scales, and has become an essential technique for monitoring AGB.

Remote sensing data sources used for the prediction of AGB include optical remote sensing, microwave radar, LiDAR technology, and terrestrial laser scanning [11–14]. They have been widely used for plant biomass monitoring in the desert regions using satellite data with coarse or medium resolution, such as Moderate Resolution Imaging Spectroradiometer [15], Landsat 8 OLI [16], and Sentinel-2 [17]. A wide range of image coverage and simplicity of data acquisition would make the remote sensing widely used in a variety of scales. Considering the sparse and dwarf distribution of vegetation in the desert, the use of satellite data with low and medium resolutions would result in a significant uncertainty and bias in the monitoring of AGB at a large scale [18,19]. Even though the high-accuracy satellite imagery avoids such issues, it would be difficult to obtain the images. For the preparation of the vmarker characterization of low height shrubs in the desert, remote sensing satellite images cannot provide the entire vertical canopy information.

Due to an active remote sensing technology used, LiDAR data has a higher spatial resolution [20] and its laser pulses can penetrate deeply into the canopy layer. This can provide accurate information regarding desert vegetation phenotypic characteristics, particularly vegetation height and canopy volume, which are essential for estimating biomass [21]. The elevation, density and intensity features obtained from 3D point clouds provide reliable parameters for the prediction of AGB [22]. Elevation reflects the height characteristics of plant communities in various ways [23] and intensity variables reflect the intensity of LiDAR pulse echoes generated at a particular location by assessing height characteristics of plant communities. These data can be used for a variety of purposes including tree classification [23], forest type classification [24], forest characterization and urban ringing [25].

Traditional biomass estimation relies on the parametric regression method, which is easy to use and has a straightforward interpretability [26]. The parameters are estimated by fitting biomass data according to multiple theoretical models. The parametric regression does not ensure the absence of a better model to form outside the range of candidate models, and the least squares method used for parameter estimation requires the existence of a normal distribution of the dependent variable or errors [27,28]. In contrast, the nonparametric methods or machine learning techniques (e.g., Random Forest, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting Machine and so on) effectively avoid such problems. It does not require a fixed model form, as there is no excessive restriction on the variables, and can be used to estimate forest biomass quantitatively using a wide range of variables [29]. The machine learning techniques are widely used for the prediction of forest growth and yield [30,31], forest site quality assessment [32,33], forest biomass prediction and so on. Several studies show no best modelling biomass technique has existed, but depending on the scope and purpose of the investigation, some techniques would likely be more suitable than others [29].

With the ability to collect information at regional and global scales, remote sensing is the efficient method that can be used to estimate forest biomass over the large areas at

the same time. At present, machine learning techniques and airborne LiDAR data have been commonly used to predict forest biomass, but research on the shrub AGB in the desert region is lacking. We constructed the AGB model applicable to the desert shrub community AGB prediction based on the airborne LiDAR data and field survey data. We evaluated both the parametric and non-parametric modeling approaches for predicting the AGB of the shrub communities in the desert region. The main objectives of our study are to: (1) identify the applicability of elevation, density and intensity variables for predicting the shrub's AGB in desert based on LiDAR data, (2) establish the nonlinear mixed-effects AGB model with the random effects at sample plot level included, and (3) compare the performance of parametric and non-parametric methods for predicting the AGB of the desert shrub community and select the most accurate and efficient model method.

2. Materials and Methods

2.1. Study Area

The study area is located in Dengkou in the northeastern part of the Ulan Buh Desert (106°38′42″E–106°57′00″E, 40°17′24″N–40°28′36″N) (Figure 1). The elevation ranges from 1048 to 1053 m above mean sea level. This area belongs to the mid-temperate arid climate zone; the average annual temperature of 7 °C~8 °C, average annual precipitation of 102~140 mm. The soil is composed of irrigated silt soil, gray desert soil, saline soil, aeolian sandy soil and light brown calcium soil. *Haloxylon ammodendron*, *Artemisia sphaerocephala* and *Nitraria* are the dominant vegetation in the area. The eastern edge of the Ulan Buh Desert is the dividing line between the wasteland and steppe in central Asia, and it is also a very important dividing line of the plant geography.

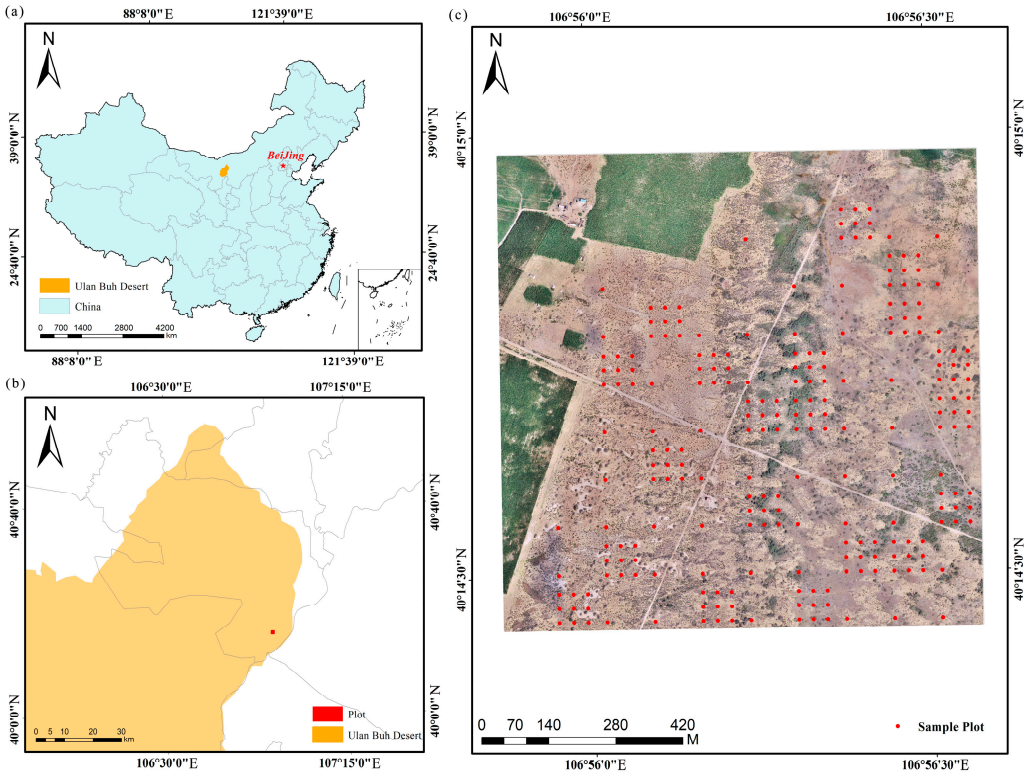

Figure 1. (**a**) Location of the study area: Ulan Buh Desert, Deng Kou in the NeiMonggol, China; (**b**) positions of permanent plots in Ulan Buh Desert; (**c**) distribution location of sample plots of 30 m × 30 m.

2.2. Data Acquisition and Preprocessing

2.2.1. Field Survey

Twenty permanent sample plots (PSPs) of 100 m × 100 m and forty-eight PSPs of 30 m × 30 m were established to represent the clustered vegetation types in July through September 2019. To facilitate subsequent calculations, the southwest corner of each 100 m × 100 m PSPs was used as a benchmark to divide the plot into nine subplots of 30 m × 30 m. The spatial distributions of 228 PSPs are shown in Figure 1. The shrub species found in the study area are *Artemisia desertorum*, *Nitraria tangutorum*, *Haloxylon ammodendron*, *Caragana korshinskii*, *Hedysarum scoparium*, *Tamarix ramosissima*, *Elaeagnus angustifolia*. Within each of the PSPs, all the standing live shrubs were measured for basal diameter (BD), total height (H) and crown width in two perpendicular directions (C1, C2). The location of the shrub community was measured using the Real-time Kine Matic (RTK) system. The above ground biomass of the shrubs were obtained based on the allometric equation of shrubs in the area (Table 1) [34]. The AGB for 228 PSPs was obtained by adding up the plot-level biomass that ranged from 0.0054 ton ha^{-1} (0.54 g m^{-2}) to 5.428 ton ha^{-1} (542.80 g m^{-2}). The average AGB of 228 PSPs was 1.1421 ton ha^{-1} (114.21 g m^{-2}) and the standard deviation was 1.0646 ton ha^{-1} (106.46 g m^{-2}).

Table 1. Allometric equation with known parameters of the shrub's AGB.

Plant Species	Formula	Allometric Exponent
Artemisia desertorum	$AGB = 393.985 * V^{0.951}$	1.135
Nitraria tangutorum	$AGB = 131.268 * V^{0.611}$	0.945
Haloxylon ammodendron	$AGB = 688.379 * V^{0.832}$	1.007
Caragana korshinskii	$AGB = 414.792 * V^{1.130}$	1.014
Hedysarum scoparium	$AGB = 170.439 * S^{1.474}$	0.816
Tamarix ramosissima	$AGB = 533.087 * V^{0.745}$	0.869
Elaeagnus angustifolia	$AGB = 317.905 * V^{1.013}$	1.071

S is the shrub's crown area by $S = \pi \times \left(\frac{C1+C2}{4}\right)^2$, V is the crown volume by $V = H \times S$.

2.2.2. LiDAR Data and Preprocessing

Airborne LiDAR data were collected by the CHCNAV AS-1300HL system with laser scanner—Riegl VUX-1LR, which was produced by RIEGL company in Niederosterreich, Austria. The average flight height was 200 m, and the average flight speed was 10 m s^{-1}. The scanner zenith angle was −33°~33°and the point cloud lateral overlap rate was 50%.

The original point cloud data was obtained by Co-pre (Figure 2). Extreme outliers were removed by the spatial distribution algorithm. The improved progressive Triangulated Irregular Network (TIN) densification was used for ground point classification [35]. The laser point cloud data were normalized to remove the effects of the unleveled ground surface by subtracting the elevation of the ground point from the elevation of the other point. The point cloud density in the area was 68.6 pts m^{-2} after preprocessing.

Three groups of indices were calculated from LiDAR point cloud data: metrics based on the elevation, density and intensity value of the points. We calculated 97 variables using the Lidar 360 Tools to identify the applicability of variables for predicting the AGB of shrubs in desert based on LiDAR data (Table 2).

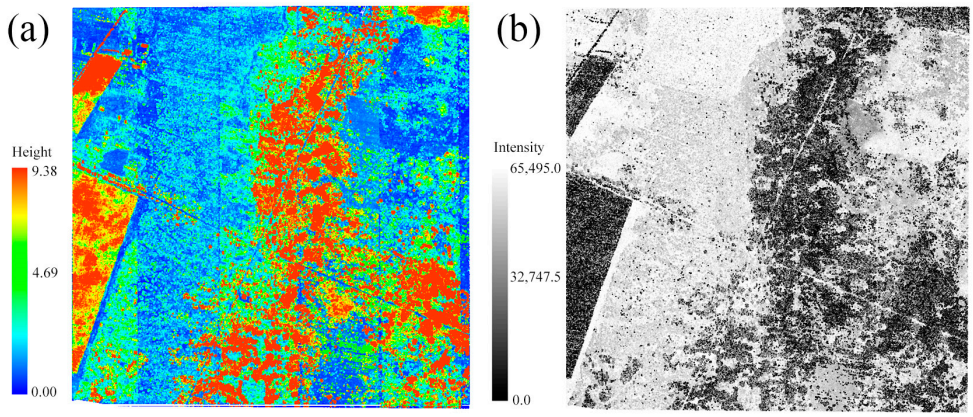

Figure 2. LiDAR point cloud pretreatment process ((**a**) refers to the normalized point cloud of sample plot; (**b**) refers to the intensity of point cloud).

Table 2. Variables calculated based on LiDAR point data.

Variables	Description	Variables	Description
CEP1, CEP5, CEP10, . . . , CEP90, CEP95, CEP99	Cumulative elevation percentile of 15 different statistical units	EMax, EMean, EMed, EStd, EVar, IMax, IMin, IMean, IMed, IStd, IVar	Maximum, Minimum, Mean, Median, Standard Deviation and Variance of Elevation and Intensity
CIP1, CIP5, CIP10, . . . , CIP90, CIP95, CIP99	Cumulative intensity percentile of 15 different statistical units	EP1, EP5, EP10 . . . , EP95, EP99	Elevation percentile of 15 different statistical units
IP1, IP5, IP10, . . . , IP90, IP95, IP99	Intensity percentile of 15 different statistical units	DM1, DM2 . . . , DM10	Point cloud density of 10 different statistical units
CEPIq, EIq, IIq	Percentile quartile spacing of Cumulative Elevation, Elevation and Intensity	EMAD, IMAD	Median absolute deviation of the median Elevation and Intensity
ECM, ESM	Mean elevation of Cube and Sqrt	HAd, IAd	Average absolute deviation of Elevation and Intensity
ECv, ICv	Coefficient of variation of Elevation and Intensity	CFd	Canopy fluctuation rate
EKu, IKu	Elevation kurtosis and intensity kurtosis	ESk, ISk	Elevation Skewness and Intensity Skewness

2.3. Parametric Regression Models

2.3.1. Base Model

Four commonly used candidate models were considered in this study as base models, and they are the Linear model (Equation (1)), Logistic model (Equation (2)), Exponential model (Equation (3)) and Richards model (Equation (4)) to fit data. The LiDAR variables were screened by a stepwise regression and VIF (Variance Inflation Factor, VIF < 5) collinearity test was performed to avoid interdependency among the multiple predictor variables. The non-linear mixed-effects (NLME) AGB model was also developed using the best performing base model. Min-Max Normalization of each variable was used to promote the convergence of the models.

$$AGB = \beta_1 + \beta_2 x_1 + \beta_3 x_2 \cdots \beta_n x_m + \varepsilon \quad (1)$$

$$AGB = \beta_1 / [1 + \beta_2 \exp(-\beta_3 x_1 - \beta_4 x_2 \cdots - \beta_n x_m)] + \varepsilon \quad (2)$$

$$AGB = \beta_1 \exp(-\beta_2 x_1 - \beta_3 x_2 \cdots - \beta_n x_m) + \varepsilon \quad (3)$$

$$AGB = \beta_1 [1 - \exp(-\beta_2 x_1 - \beta_3 x_2 \cdots - \beta_n x_m)] + \varepsilon \quad (4)$$

where x_m is the stand parameter extracted by LiDAR; β_1, β_2, β_3, β_4, β_n are parameters to be estimated, and ε is an error term.

The best basic model was selected based on the following statistical criteria:

$$\bar{e} = \sum e_t / N = \sum_{t=1}^{N} \left(AGB_t - A\hat{G}B_t \right) / N \quad (5)$$

$$\sigma^2 = \sum_{t=1}^{N} (e_t - \bar{e})^2 / (N-1) \quad (6)$$

$$TRE = 100 * \sum_{t=1}^{N} \left(AGB_t - A\hat{G}B_t \right)^2 / \sum_{t=1}^{N} (AGB_t)^2 \quad (7)$$

$$R^2 = 1 - \sum_{t=1}^{N} \left(AGB_t - A\hat{G}B_t \right)^2 / \sum_{t=1}^{N} \left(AGB_t - \overline{AGB} \right)^2 \quad (8)$$

$$RMSE = \sqrt{\bar{e}^2 + \sigma^2} \quad (9)$$

where AGB_t and $A\hat{G}B_t$ are the aboveground biomasses estimated by the allometric equation and predicted by the newly developed AGB model, respectively, and $\overline{AGB}$ is the mean aboveground biomass by the allometric equation; and N is the number of sample plots; and $\bar{e}$, σ^2, R^2, and *RMSE* are the mean bias, variance of bias, coefficient of determination, and root mean square error, respectively. *RMSE* is defined as the combination of the mean bias and its variance and is the most important evaluation criterion of the model.

2.3.2. Nonlinear Mixed-Effects Model

The mixed-effects model is based on the regression function on the fixed-effects parameters and the random-effects parameters [36]. The general form of the single-level mixed-effects model is:

$$\begin{cases} y_{ij} = f(\boldsymbol{\phi}_i, X_{ij}) + \varepsilon_{ij}, i = 1, \ldots, M, j = 1, \ldots, n_i, \\ \boldsymbol{\phi}_i = \mathbf{A}_i \boldsymbol{\beta} + \sum\limits_{k=1}^{K} \mathbf{B}_i^{(k)} \mathbf{u}_i^{(k)}, \\ \mathbf{u}_i^{(k)} \sim N(0, \boldsymbol{\psi}^{(k)}), \mathrm{cov}(\mathbf{u}_i^{(k)}, \mathbf{u}_i^{(l)}) = 0, k \neq l, k = 1, \ldots, K, l = 1, \ldots K, \\ \boldsymbol{\varepsilon}_i \sim N(0, \mathbf{R}_i), \boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^T \end{cases} \quad (10)$$

where y_{ij} is the jth observation on the ith subject for a dependent variable, M is the number of subjects, n_i is the number of observations on the ith subject, $f(.)$ is a real-valued and differentiable nonlinear function of a $p \times 1$ vector of the subject-specific parameters $\boldsymbol{\phi}_i$ and a $s \times 1$ covariate vector X_{ij}, β is a $p0 \times 1$ vector of fixed effects, $\mathbf{u}_i^{(k)}$ is a $q^{(k)} \times 1$ vector of the random effects assumed to be normally distributed with a mean of zero and a variance-covariance matrix $\boldsymbol{\psi}^{(k)}$, and $\mathbf{A}_i$ and $\mathbf{B}_i^{(k)}$ are design matrices. For different k, $l \in \Omega$ $(k \neq l)$, $\mathbf{u}_i^{(k)}$ and $\mathbf{u}_i^{(l)}$ are independent of each other. The error term $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})$ is assumed to normally distributed with a mean of zeros and a covariance matrix of $\mathbf{R}_i$, independent of $\mathbf{u}_i^{(k)}$ s. More detailed explanation of nonlinear mixed-effects modeling can be found in Fu and Tang [37].

The shrub vegetation cover in the sample plots was used as a random effect factor and is classified into three levels based on related studies: 0–40%, 40–60% and 60–100%. The optimal NLME AGB model was determined by testing the combination of random-

effect factors using AIC (Akaike information criterion, Equation (10)) and BIC (Bayesian information criteria, Equation (11)). As both AIC and BIC give almost similar values, the former, which is the most common one, can be used, rather than both. When the AIC of any model is ≥2 than others, then former model can be considered significantly better than that model [38]. The NLME model was implemented based on the 'nlme' package in R-4.2.

$$AIC = 2k - \ln(L) \tag{11}$$

$$BIC = k\ln(N) - 2\ln(L) \tag{12}$$

where k is the number of model parameters, N is the number of samples, L is the likelihood function value.

2.4. Machine Learning Algorithms

2.4.1. Random Forest

Random Forest (RF) uses the majority voting method and the final results are derived by combining the decision results of each tree [39]. The results can be expressed as:

$$H(x) = arg\mathbf{max}_{\gamma}\sum_{i=1}^{k} I(h_i(x) = Y) \tag{13}$$

where $H(x)$ is the combinatorial model of classification or regression, representing the results of RF. $h_i(x)$ is the single model of classification or regression, representing the result of CART (Classification and Regression Tree). Y is the output variable. $I(\cdot)$ is the indicator function.

Key parameters of the random forest algorithm are the number of classification trees and the number of node features. After parameter tuning, the Ntree (Number of decision tree) and Mtry (Number of variables in the decision tree) values were 1000 and 13, respectively. The RF classifier is implemented based on the 'randomForest' [40] package in R-4.2.

2.4.2. Support Vector Machine

A Support Vector Machine (SVM) has four main types of the kernel functions: Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid Kernels [41]. In our study, we applied the RBF kernel for the AGB prediction, which has the optimization result in applications [42]. The equation for RBF is:

$$K(x, x_i) = \mathbf{exp}(-\gamma \parallel x - x_i \parallel^2), \gamma > 0 \tag{14}$$

where x is the feature vector of the recognition sample; x_i is the feature vector of the training sample; γ is the parameter controlling the width of the Gaussian kernel.

SVM was modeled using the "e1071" package in R-4.2 to use the grid search method to identify the optimal Gamma and Cost parameter combinations [43] The model accuracy is the highest when Gamma is taken as 0.001 and Cost is 13.

2.4.3. K-Nearest-Neighbor

The traditional K-Nearest-Neighbor (KNN) algorithm was used to calculate distance and similarities among all the labeled instances in the training set for each test instance [44]. Normally, Euclidean distance was used to measure distance. The Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{15}$$

where (x_i, y_i) are the coordinates of n points.

Using a comprehensive method, this study determines the optimal k value for the KNN algorithm based on the exhaustive method. With a k value of 14, the accuracy of

the model is optimal. The KNN model was constructed based on the 'class' package in R-4.2 [45].

2.4.4. Gradient Boosting Machine

The Gradient Boosting Machine (GBM) is based on the principle of constructing new base-learners relevant to the negative gradient of the loss function and associated with the whole ensemble [46]. The goal is to obtain an estimate or approximation of $\hat{F}(\mathbf{x})$, of the function $F^*(\mathbf{x})$ mapping $\mathbf{x}$ on to y, that minimizes the expected value of some specified loss function $L(y, F(\mathbf{x}))$ over the joint distribution of all the $(y, \mathbf{x})$ values,

$$F^* = \underset{F}{\operatorname{argmin}} E_{y,\mathbf{x}} L(y, F(\mathbf{x})) = \underset{F}{\operatorname{argmin}} \underbrace{E_{\mathbf{x}}[\overbrace{E_y(L(y, F(\mathbf{x})))}^{\text{expected y loss}} | \mathbf{x}]}_{\text{expectation over the whole dataset}} \tag{16}$$

In general, the choice of the loss function is up to the researcher and there are not only a variety of loss functions, but also the possibility of implementing one's own task-specific loss [46]. A classic loss function, which is commonly used is the squared-error L2 loss is: $\Psi(y, f)_{L_1} = \left| y - f \right|$. The parameters n.trees, interaction.depth and shrinkage were set as 4000, 1 and 0.001, respectively. The GBM algorithm was modeled based on the 'gbm' package in R-4.2 [47].

2.4.5. Multivariate Adaptive Regression Splines Model

Multivariate Adaptive Regression Splines (MARS) uses the tensor product of spline function as the basic function and is divided into three steps: forward process, backward pruning process and model selection. The MARS segments the data by adaptively selecting nodes and generating the corresponding basic functions, and finally constructs the model by adding basic functions, and the expression is:

$$f(x) = \alpha_0 + \sum_{i=1}^{N} \alpha_i H_i(x) \tag{17}$$

where $f(x)$ is the predicted value of the target variable; α_0 is the intercept; α_i is the coefficient corresponding to the i-th basic function; $H_i(x)$ is the i-th basic function; N is the number of basic functions. The basic function can be expressed as:

$$H_i(x) = \begin{cases} \max(0, E - x) \\ \max(0, x - E) \end{cases} \tag{18}$$

where E is the threshold value of the input variable; x is the predictor variable. Two hyper-parameters "degree" and "nprune" need to be set in the modeled MARS. The default value "1" of "degree" indicates that there is no interaction between independent variables, and "nprune" represents the maximum number of items in the model, and the value was determined as 12 by the test. The MARS model was constructed based on the 'caret' package in R-4.2 [48].

2.5. Model Evaluation

In this study, the data set was divided into training and validation sets according to the proportion of 7:3. Seventy percent of data was used for modeling and 30% for validation. Both the basic model and the NLME model were evaluated by an independent dataset. The predicted and observed AGB values were used to calculate prediction statistics (R^2, RMSE and TRE) using Equations (5)–(8). Based on the best model, utilizing the DSA (Data-based Sensitivity Analysis) method, the importance of variables was explored.

3. Results

3.1. Parametric Regression Models

3.1.1. Base Models

The Pearson correlation test showed that most LiDAR variables had high correlations and positive relationships with shrub biomass. Fifteen predictor variables were obtained by stepwise regression screening, and CEP30 (Cumulative Elevation Percentile of 30%), EP40 (Elevation Percentile of 40%), ESk (Elevation Skewness), CIP1(Cumulative Intensity percentile of 1%), and CIP95 (Cumulative Intensity Percentile of 95%) variables were then selected from subsequent model fitting with the use of VIF (Variance Inflation Factor, VIF < 5) as a collinearity test method(Figure 3).

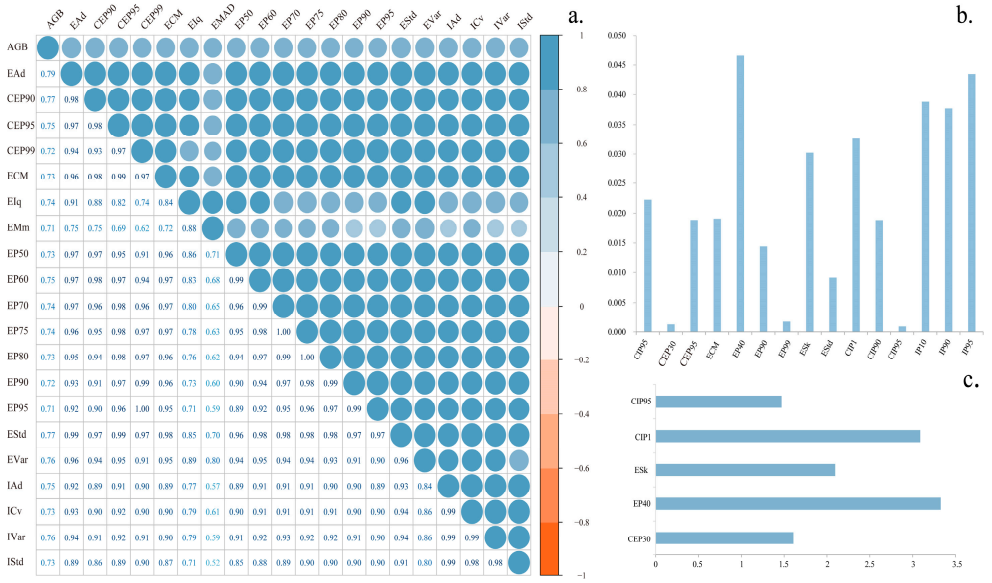

Figure 3. Correlation coefficients of the variables and screening index ((**a**) the variables with the top 20 correlation coefficients; (**b**) fifteen significant variables in a stepwise regression analysis, *y*-axis is the p value; (**c**) *x*-axis is the VIF value).

The optimal inversion basic model was identified from the four biomass candidate models based on the training data set. There was a similar level of accuracy among the four basic models on the training set, with the logistic model having the best accuracy (Table 3). The validation results, which were calculated using 30% data, showed the best accuracy of the Exponential model (R^2 = 0.6169 and *TRE* = 22.7290). Even though the training accuracy of the Exponential model was slightly inferior to that of the Logistic model, which has six parameters, the former model is simpler than the latter model, and the AIC difference < 2 was therefore selected as a basis for developing the NLME AGB prediction model.

Table 3. Fit statistics of four candidate basic models. (RMSE, root mean square error, units were g m^{-2}; R^2, coefficient of determination; TRE, total system error; AIC, Akaike's information criterion; BIC, Bayesian Information Criterion).

Model	Number of Parameters	Training Data					Validation Data		
		RMSE	R^2	TRE	AIC	BIC	RMSE	R^2	TRE
Linear	6	60.9266	0.6362	20.2317	1771.09	1792.58	79.8046	0.5467	25.7699
Richards	6	57.2373	0.6789	17.4414	1751.23	1772.71	85.4208	0.4807	25.1191
Logistic	7	55.9541	0.6931	16.5402	1746.02	1770.57	74.8961	0.6006	23.2180
Exponential	6	56.5880	0.6861	16.9811	1747.60	1769.09	73.3495	0.6169	22.7290

3.1.2. NLME Models

Considering the six parameters in the model (β_1~β_6), there were 63 different combinations of the random effects for the basic Exponential model. Forty-two of all NLME model variants converged with the meaningful parameter estimates, the smallest AIC was 1729.73 (combination of the parameter 4 and 5, Figure 4).

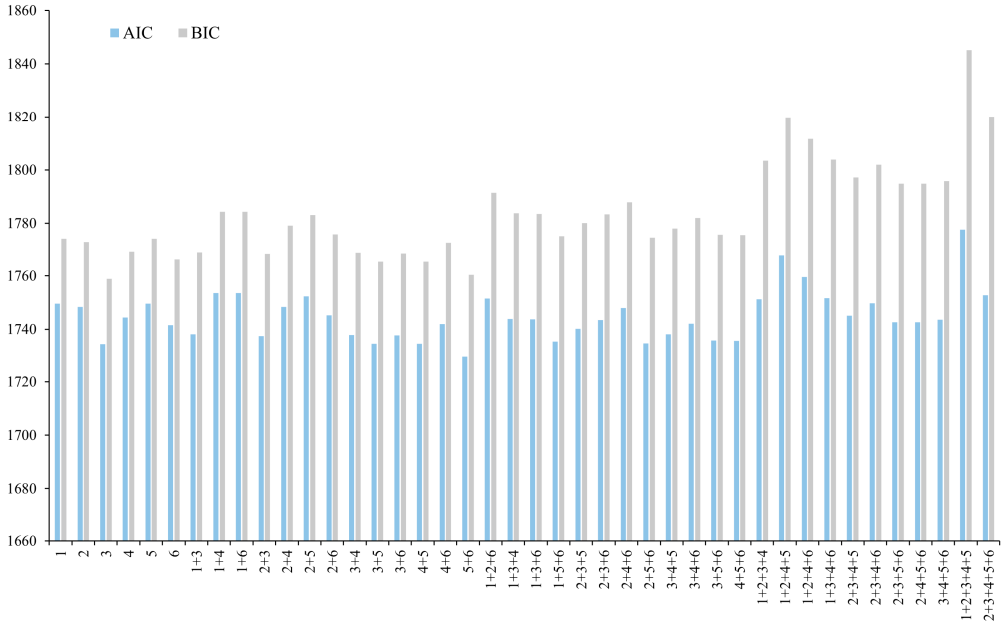

Figure 4. Forty-two combinations of the random effects for the basic model 4. (1~6 refer to the parameters needs to be included as random effects; + refer to the combinations of the random effects).

Through the performance of the models in the training set and validation set, the following NLME models had optimal fit statistics (training data: R^2 = 0.7509, *RMSE* = 50.4077, *TRE* = 13.0180; validation data: R^2 = 0.6977, *RMSE* = 65.1613, *TRE* = 15.6269).

$$AGB = \beta_1 \exp\left[\begin{array}{c} -\beta_2 * CEP30 - \beta_3 * EP40 - \beta_4 * ESk \\ -(\beta_5 + \mu_{5i}) * CIP1 - (\beta_6 + \mu_{6i}) * CIP95 \end{array}\right] + \varepsilon$$

where *AGB* is the aboveground biomass; β_1~β_6 are the fixed effects parameters; *CEP30*, *EP40*, *ESk*, *CIP1* and *CIP95* refer to the cumulative elevation percentile of 30%, elevation percentile of 40%, elevation skewness, cumulative intensity percentile of 1%, cumulative intensity percentile of 95%; μ_{5i} and μ_{6i} are the random effects due to the shrub's coverage of the sample plots on β_5 and β_6, respectively.

3.1.3. Parameter Estimates

Most of the parameter estimates were significantly different from zero ($p < 0.05$), and their magnitudes and signs could meet biological logics (Table 4). The NLME AGB model is:

$$AGB = 24.436 * \exp\left[\begin{array}{c} 0.106 * CEP30 + 0.928 * EP40 + 0.462 * ESk \\ -(2.197 + 1.062) * CIP1 - (-2.272 + 0.586) * CIP95 \end{array}\right] + \varepsilon$$

where *AGB* is the aboveground biomass; *CEP30*, *EP40*, *Esk*, *CIP1* and *CIP95* are the parameters obtained by the airborne LiDAR: cumulative elevation percentile of 30%, elevation percentile of 40%, elevation skewness, cumulative intensity percentile of 1%, cumulative intensity percentile of 95%, respectively.

Table 4. Parameter estimates of the nonlinear regression model including the NLME AGB model.

Parameter	Variables	Liner	Logistic	Richards	Exponential	NLME
β_1	/	−60.340	−182.121	−20.556	26.834	24.436
β_2	*CEP30*	−29.370	−3.922	−0.336	−0.361	−0.106
β_3	*EP40*	270.500	0.188	−1.059	−0.843	−0.928
β_4	*Eske*	86.700	0.206	−0.834	−0.247	−0.462
β_5	*CIP30*	−144.790	−0.102	1.901	2.555	2.197
β_6	*CIP95*	192.47	−1.572	−2.327	−2.434	−2.272
β_7	/	/	1.117	/	/	/
μ_{5i}	/	/	/	/	/	1.062
μ_{6i}	/	/	/	/	/	0.586

"/" means that this parameter term does not exist in the model.

3.2. Machine Learning

Five machine learning algorithms were used to construct the AGB model for shrub communities in the desert region. Based on the model training data, the SVM-model achieved the best accuracy with the lowest values of RMSE and TRE, and the highest R^2, followed by Earth, KNN and GBM-models, respectively (Table 5) where RF-model scores were lower for these indicators. Similar trends can be observed by comparing the predictive performance of all the models in validation data. The SVM-model still scored the best (R^2 = 0.8962, *RMSE* = 38.1919, *TRE* = 5.4063). However, the performance of the RF-model was better than that of the MARS and GBM-model in the validation dataset, and the accuracy of GBM model was the lowest.

Table 5. Fit statistics of machine learning models. (RF, random forest; SVM, support vector machine; KNN, K-Nearest Neighbor; MARS, Multivariate Adaptive Regression Splines; GBM, Gradient Boosting Machine; Units of RMSE is g m^{-2}).

Model	Training Data			Validation Data		
	RMSE	R^2	TRE	RMSE	R^2	TRE (g m^{-2})
RF	49.7021	0.7579	13.0055	62.8252	0.7191	14.4489
SVM	26.8635	0.9293	3.5050	38.1919	0.8962	5.4063
KNN	46.4816	0.7882	11.3961	63.5297	0.7128	16.5406
MARS	32.0995	0.8990	4.8997	60.8154	0.7367	10.4576
GBM	48.8357	0.7662	13.5585	74.0015	0.6103	23.2776

3.3. Model Evaluation

The performance of all the models were evaluated by an independent data set using three statistical indicators (RMSE, R^2 and TRE). It was found that the machine learning algorithms had advantages over the traditional parametric model (Tables 1 and 3). The R^2 of the best SVM-model increased by 20% and the TRE and RMSE reduced by 10% when compared with the best traditional NLME model. At the same time, the stability of the SVM-model in both the fitting data and validation data were much better than the NLME model. Among the random distributions of the residuals produced, the SVM model and the MARS model more effectively reduce the heteroscedasticity than other models (Figure 5). The residuals of the six models were concentrated towards the zero line when biomass was smaller, and residuals were relatively spread away from the zero line when biomass increased. Figure 6 shows the AGB map of approximately one square kilometer drawn based on the SVM model.

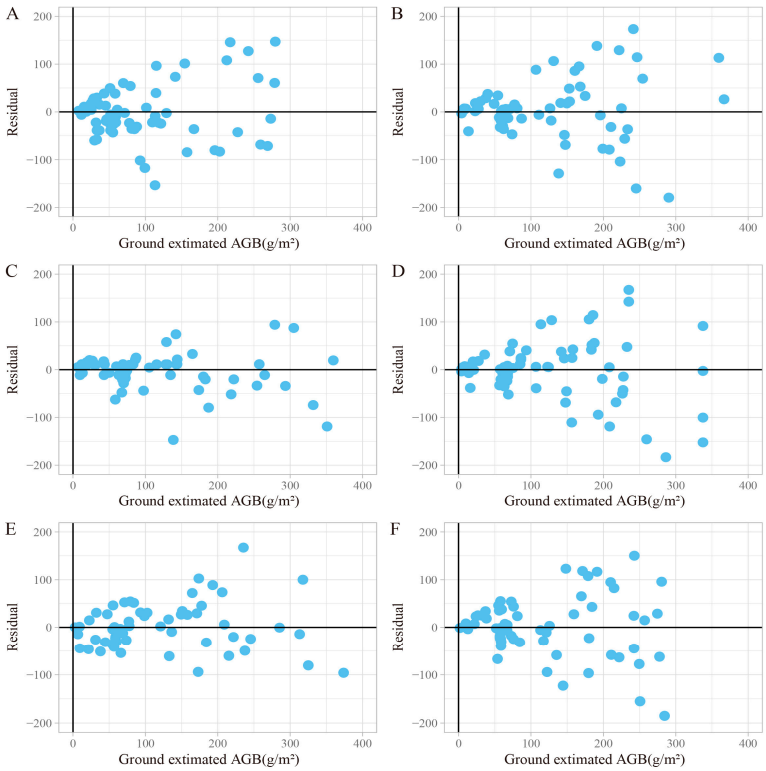

Figure 5. Predicted ground estimated AGB residuals distribution of six models (**A**–**F**) refer to the NLME model, Random Forest model, Support vector machine model, K-Nearest Neighbor, MARS model and the GBM model.

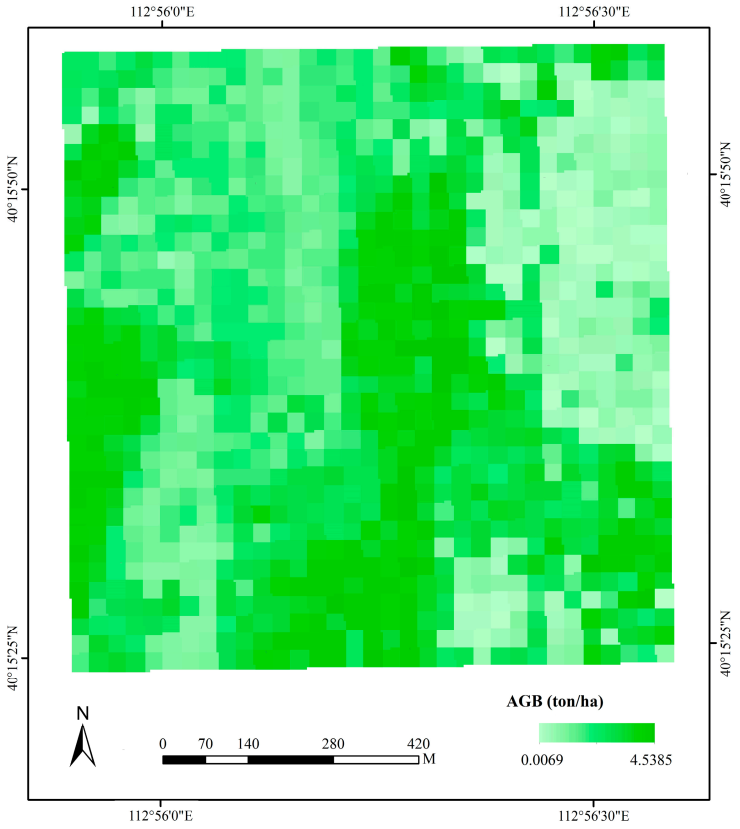

Figure 6. AGB map based on the SVM model at a sample plot as an example.

3.4. Variables Importance

The SVM-model achieved the best accuracy, and the ten most important variables of DSA results were Eku, Emea, Iku, EMAD, DM5, DM4, Isk, CEP5, DM6 and DM10 (Figure 7). Among them, Eku was the flatness of the height distribution of all the points in the statistical unit and had the highest relative importance. Emea and EMAD also had a high relative importance. These elevation variables were critical for the inversion of biomass. In the cumulative height percentage, CEP5 was relatively high. Among the intensity variables, the Iku and Isk of the response intensity appeared more important, while the variables related to Intensity percentile and cumulative intensity percentage were generally less important. Four of the ten density variables were ranked in the top ten, DM4 and DM5 in the mid-height slice were more important than others.

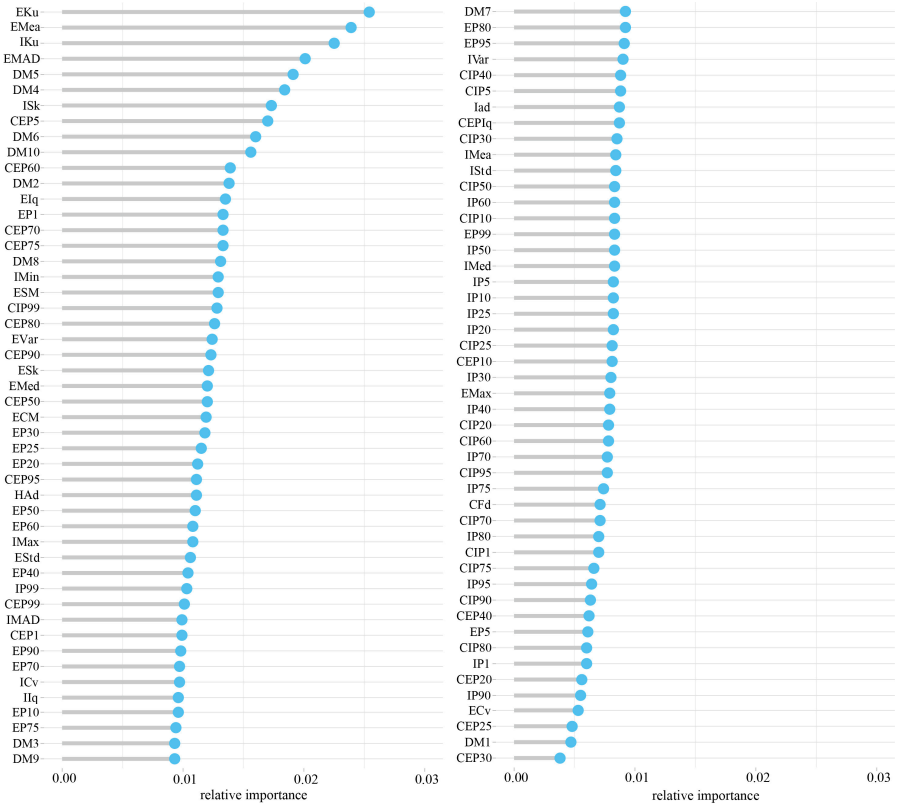

Figure 7. Lollipop plot with the importance of DSA (Data-based Sensitivity Analysis) variables for the SVM-Model.

4. Discussion

Based on the airborne LiDAR data, a shrub biomass prediction model for a sample plot-level in the desert was constructed by utilizing both the parametric regression and machine learning methods. This study indicates that the best prediction model (R^2 = 0.8962) can accurately predict shrub community biomass at the plot level. Modeling approaches employed to different forest biomass modeling studies might be largely different, the number of predictor variables in the models used might be different, the amount of data or number of observations used in modeling might be also different, and so on. Thus, comparing any previously developed forest biomass models against our shrub biomass model might not be relevant. However, to some extent, it would be useful if compared. In most studies, vegetation biomass dynamics in the arid regions were monitored based directly on the ground survey data [15,16,49–51]. However, satellite-based biomass models were observed to be relatively inaccurate. Liang [49] found that the NDVI-based AGB model performed the best amongst all the single-factor models for the MODIS satellite (R^2 = 0.58), similar to Jin's study [52] of the desert grassland of northern China (R^2 = 0.45~0.50). The

accuracy of the satellite prediction model depends on the number and representativeness of the samples, as well as the spatial consistency between the satellite pixel sizes and ground sampling area [2]. Compared to remote sensing data, LiDAR provides a more accurate reflection of shrub characteristics in three dimensions, which is important for the estimation of AGB [53]. LiDAR can provide more accurate information about the vertical structure of shrub communities compared to the point clouds produced by UAV photogrammetry [54]. Based on 35 statistical variables in ALS data, Li used the random forest regression model and a stepwise multiple regression model to estimate shrub biomass. The results indicated that two models can explain >74% of changes in shrub biomass [55]. In comparison with relevant research, our study had a better fitting accuracy (R^2 = 0.8962), indicating that the application of LiDAR can significantly improve AGB estimation in the desert region.

For the prediction of the shrub's AGB in the desert region, machine learning algorithms have certain advantages over the traditional regression methods. The previous studies have indicated that the machine learning models can be more effective for ecosystem research than the parametric regression models due to their complex structures and their ability to accurately capture the nonlinear relationships between the variables [56–58]. A key feature of the machine learning algorithms is that they effectively avoid the problems with variable screening and multiple collinearities, and do not require data to conform the certain distribution characteristics [59,60]. Screening variables is necessary in order to reduce the complexity of the model, and finding the initial values of the model parameters could be time-consuming when using the parametric regression [61]. The machine learning algorithm eliminates these steps, simply adjusts model parameters, uses more diverse input variables, and can be used to analyze the multiple noisy data sets simultaneously [62]. These characteristics are more conducive to understanding by local managers. One of the most common criticisms of machine learning is its black box nature, which refers to the challenge in understanding how algorithms make their decision [63]. However, it is important to note that every algorithm has an internal logic, and its black box nature is only apparent for those who use this [64]. Many studies are dedicated to explaining the black box of machine learning, which is crucial for its widespread application [65]. The accuracy of the machine learning model improved by 10% in our study, but some researchers have indicated that there is no universal evaluation standard between machine learning algorithms and traditional methods, and as such, studies typically focus on individual cases [66,67]. A consideration should be made regarding the compatibility of the models in traditional parameter estimation methods as well as how to analyze measurement errors in machine learning.

Each machine learning algorithm has its own advantages in handling data of a variety of dimensions [68]. Among the five machine learning algorithms used, SVM performed the best fitting performance and the optimal residual distribution (Table 5), which is an indication of its ability to deal with the interactions between nonlinear features of variables and its strong generalization capability [69]. A study by Safari et al. [70] compared nine methods of mapping AGB based on optical satellite data, and reported that SVM was more accurate in terms of the coefficient of determination. The SVM algorithm searches for the optimal hyperplane to minimize training errors and suitably generalizes a given model with limited training samples, but it suffers from the parameter allocation problems that seriously affect the results [71]. In our study, RF, KNN and MARS perform similarly and the feature of RF with built-in generalization error estimation makes it the most stable among other models in the training and validation data sets. The performance of GBM in this study is similar to that of the RF in the training set, as both the algorithms are boosting algorithms, which can handle various types of data flexibly. Five common machine learning algorithms (RF, SVM, KNN, MARS, and GBM) were systematically compared, but each has disadvantages to a varying degree, and thus, it was meaningful to explore the integrated modeling methods by combining advantages of the multiple algorithms [72,73].

The previous studies conducted in analogous ecosystems showed that height [55], volume [74,75], or the approximation of volume [76] (e.g., the product of the basal area

and height) to be a strong proxy of shrub biomass. The DSA results showed that Eku was the most influential variable. Eku refers to the flatness of the elevation distribution across all the points within the statistical unit, which can reflect the elevation of the point, the average height in the unit, the total number of points, and the standard deviation of the point cloud elevation distribution. Variables associated with elevation, such as the Emea, EMAD, and Elevation Percentage, generally exhibited a greater importance. A study by Li et al. [55] underscored the significance of variables linked to elevation changes in all the random forest models. The intensity variables, which are partially based on the vegetation reflectivity [77], demonstrated relatively lower importance, although it is worth mentioning that Iku, encompassing information on the total points and average elevation of the points, ranked as the third most important variable. Density variables exhibited great importance, as they play a crucial role in predicting AGB. Notably, the importance ranking of DM1 was considerably low, likely attributable to the fact that density variables were divided into ten height-based slices, with the lowest slice failing to adequately capture volume-related information of the shrub community. In summary, while variable importance may differ across studies, the majority of variables related to the shrub's height and volume would be more important than other variables.

The rapid advancement of remote sensing technology and the popularity of machine learning algorithms have led to more efficient methods for estimating AGB of desert shrub community [68,78]. Due to the heterogeneity and complexity of vegetation types, soil, and topography in the desert region, combining LiDAR data with machine learning algorithms could be more appropriate for predicating AGB. Our method is claimed to be innovative, as this accurately capture and monitor the shrub AGB changes in the desert region. Although our study has achieved the encouraging results, further work is needed to evaluate a combination of multi-source remote sensing data and machine learning algorithms for monitoring shrub biomass dynamics in the desert region at a large scale.

5. Conclusions

We constructed the parametric model and nonparametric algorithms for prediction of the shrub's above ground biomass in the desert region using density variables, intensity variables and elevation variables, the information of which were derived from airborne LiDAR data. Our results suggest that the elevation and intensity variables can effectively predict the aboveground biomass of vegetation in the desert region at a sample plot-level. The SVM-model outperformed the nonlinear mixed-effects AGB model and reduced the heteroscedasticity more effectively. Our study shows the applicability of LiDAR in monitoring the desert shrub community biomass and the advantage of the nonparametric model for biomass inversion. The presented method and model can be used in systematic studies of carbon stock and assessments of ecological and social values in the desert region. The data that can be generated from the methods applied In our study for the biomass inversion of shrub communities in the desert, based on airborne LiDAR, will provide a strong scientific basis for desert researchers and pertinent decision makers to control and manage desertification.

Author Contributions: Conceptualization, D.X. and L.F. (Liyong Fu); methodology, D.X.; software, D.X. and L.F. (Liyong Fu); validation, D.X., L.F. (Liyong Fu) and H.H.; formal analysis, D.X., H.H. and Q.C.; investigation, D.X.; data curation, D.X. and L.F. (Linyan Feng); writing—original draft preparation, D.X. and L.F. (Liyong Fu); writing—review and editing, L.F. (Linyan Feng) and R.P.S.; visualization, Q.L., L.F. (Liyong Fu) and R.P.S.; supervision, L.F. (Liyong Fu)and R.P.S.; project administration, L.F. (Liyong Fu); funding acquisition, L.F. (Liyong Fu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the 14th Five-Year Plan Pioneering Project of High Technology Plan of the National Department of Technology (Grant No. 2021YFD2200405).

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

References

1. Grainger, A. *The Threatening Desert: Controlling Desertification*; Routledge: London, UK, 2013.
2. Mao, P.; Ding, J.; Jiang, B.; Qin, L.; Qiu, G.Y. How can UAV bridge the gap between ground and satellite observations for quantifying the biomass of desert shrub community? *ISPRS J. Photogramm. Remote Sens.* **2022**, *192*, 361–376. [CrossRef]
3. Poulter, B.; Frank, D.; Ciais, P.; Myneni, R.B.; Andela, N.; Bi, J.; Broquet, G.; Canadell, J.G.; Chevallier, F.; Liu, Y.Y. Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle. *Nature* **2014**, *509*, 600–603. [CrossRef] [PubMed]
4. Tian, Y.; Zhang, Q.; Huang, H.; Huang, Y.; Tao, J.; Zhou, G.; Zhang, Y.; Yang, Y.; Lin, J. Aboveground Biomass of Typical Invasive Mangroves and Its Distribution Patterns Using UAV-LiDAR Data in a Subtropical Estuary: Maoling River Estuary, Guangxi, China. *Ecol. Indic.* 2022, 136, p. 108694. Available online: https://www.sciencedirect.com/science/article/pii/S1470160X22001650 (accessed on 11 December 2022).
5. Perez-Quezada, J.; Delpiano, C.; Snyder, K.; Johnson, D.; Franck, N. Carbon pools in an arid shrubland in Chile under natural and afforested conditions. *J. Arid Environ.* **2011**, *75*, 29–37. [CrossRef]
6. Ahlström, A.; Raupach, M.R.; Schurgers, G.; Smith, B.; Arneth, A.; Jung, M.; Reichstein, M.; Canadell, J.G.; Friedlingstein, P.; Jain, A.K. The dominant role of semi-arid ecosystems in the trend and variability of the land CO_2 sink. *Science* **2015**, *348*, 895–899. [CrossRef]
7. Tang, Z.; Xu, W.; Zhou, G.; Bai, Y.; Li, J.; Tang, X.; Chen, D.; Liu, Q.; Ma, W.; Xiong, G. Patterns of plant carbon, nitrogen, and phosphorus concentration in relation to productivity in China's terrestrial ecosystems. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4033–4038. [CrossRef]
8. Tian, S.; Liu, X.; Jin, B.; Zhao, X. Contribution of Fine Roots to Soil Organic Carbon Accumulation in Different Desert Communities in the Sangong River Basin. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10936. [CrossRef]
9. Zhou, W.; Li, H.; Xie, L.; Nie, X.; Wang, Z.; Du, Z.; Yue, T. Remote sensing inversion of grassland aboveground biomass based on high accuracy surface modeling. *Ecol. Indic.* **2021**, *121*, 107215. [CrossRef]
10. García, M.; Riaño, D.; Chuvieco, E.; Danson, F.M. Estimating biomass carbon stocks for a Mediterranean forest in central Spain using LiDAR height and intensity data. *Remote Sens. Environ.* **2010**, *114*, 816–830. [CrossRef]
11. Fassnacht, F.; Hartig, F.; Latifi, H.; Berger, C.; Hernández, J.; Corvalán, P.; Koch, B. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sens. Environ.* **2014**, *154*, 102–114. [CrossRef]
12. Dalponte, M.; Frizzera, L.; Ørka, H.O.; Gobakken, T.; Næsset, E.; Gianelle, D. Predicting stem diameters and aboveground biomass of individual trees using remote sensing data. *Ecol. Indic.* **2018**, *85*, 367–376. [CrossRef]
13. Ahamed, T.; Tian, L.; Zhang, Y.; Ting, K. A review of remote sensing methods for biomass feedstock production. *Biomass Bioenergy* **2011**, *35*, 2455–2469. [CrossRef]
14. Lu, D.; Chen, Q.; Wang, G.; Liu, L.; Li, G.; Moran, E. A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. *Int. J. Digit. Earth* **2016**, *9*, 63–105. [CrossRef]
15. John, R.; Chen, J.; Giannico, V.; Park, H.; Xiao, J.; Shirkey, G.; Ouyang, Z.; Shao, C.; Lafortezza, R.; Qi, J. Grassland canopy cover and aboveground biomass in Mongolia and Inner Mongolia: Spatiotemporal estimates and controlling factors. *Remote Sens. Environ.* **2018**, *213*, 34–48. [CrossRef]
16. Zandler, H.; Brenning, A.; Samimi, C. Quantifying dwarf shrub biomass in an arid environment: Comparing empirical methods in a high dimensional setting. *Remote Sens. Environ.* **2015**, *158*, 140–155. [CrossRef]
17. Forkuor, G.; Zoungrana, J.-B.B.; Dimobe, K.; Ouattara, B.; Vadrevu, K.P.; Tondoh, J.E. Above-ground biomass mapping in West African dryland forest using Sentinel-1 and 2 datasets—A case study. *Remote Sens. Environ.* **2020**, *236*, 111496. [CrossRef]
18. Li, S.; Su, P.; Zhang, H.; Zhou, Z.; Xie, T.; Shi, R.; Gou, W. Distribution Patterns of Desert Plant Diversity and Relationship to Soil Properties in the Heihe River Basin, China. *Ecosphere* **2018**, *9*, e02355. Available online: https://esajournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/ecs2.2355 (accessed on 11 December 2022).
19. Avitabile, V.; Baccini, A.; Friedl, M.A.; Schmullius, C. Capabilities and limitations of Landsat and land cover data for aboveground woody biomass estimation of Uganda. *Remote Sens. Environ.* **2012**, *117*, 366–380. [CrossRef]
20. Liu, Q.; Fu, L.; Wang, G.; Li, S.; Hu, K. Improving Estimation of Forest Canopy Cover by Introducing Loss Ratio of Laser Pulses Using Airborne LiDAR. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 567–585. [CrossRef]
21. Zhou, L.; Meng, R.; Tan, Y.; Lv, Z.; Zhao, Y.; Xu, B.; Zhao, F. Comparison of UAV-based LiDAR and digital aerial photogrammetry for measuring crown-level canopy height in the urban environment. *Urban For. Urban Green.* **2022**, *69*, 127489. [CrossRef]
22. Luo, S.; Chen, J.M.; Wang, C.; Gonsamo, A.; Xi, X.; Lin, Y.; Qian, M.; Peng, D.; Nie, S.; Qin, H. Comparative Performances of Airborne LiDAR Height and Intensity Data for Leaf Area Index Estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 300–310. [CrossRef]
23. Donoghue, D.N.; Watt, P.J.; Cox, N.J.; Wilson, J. Remote sensing of species mixtures in conifer plantations using LiDAR height and intensity data. *Remote Sens. Environ.* **2007**, *110*, 509–522. [CrossRef]
24. Antonarakis, A.; Richards, K.S.; Brasington, J. Object-based land cover classification using airborne LiDAR. *Remote Sens. Environ.* **2008**, *112*, 2988–2998. [CrossRef]
25. Yunfei, B.; Guoping, L.; Chunxiang, C.; Xiaowen, L.; Hao, Z.; Qisheng, H.; Linyan, B.; Chaoyi, C. Classification of LIDAR point cloud and generation of DTM from LIDAR height and intensity data in forested area. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2008**, *37*, 313–318. [CrossRef]

26. Senthil Kumar, A.; Sudheer, K.; Jain, S.; Agarwal, P. Rainfall-runoff modelling using artificial neural networks: Comparison of network types. *Hydrol. Process. Int. J.* **2005**, *19*, 1277–1291. [CrossRef]
27. Horowitz, J.L. *Semiparametric and Nonparametric Methods in Econometrics*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 12. [CrossRef]
28. Zang, H.; Lei, X.; Zeng, W. Height–diameter equations for larch plantations in northern and northeastern China: A comparison of the mixed-effects, quantile regression and generalized additive models. *For. Int. J. For. Res.* **2016**, *89*, 434–445. [CrossRef]
29. Aertsen, W.; Kint, V.; Van Orshoven, J.; Özkan, K.; Muys, B. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecol. Model.* **2010**, *221*, 1119–1130. [CrossRef]
30. Bond-Lamberty, B.; Rocha, A.V.; Calvin, K.; Holmes, B.; Wang, C.; Goulden, M.L. Disturbance legacies and climate jointly drive tree growth and mortality in an intensively studied boreal forest. *Glob. Chang. Biol.* **2014**, *20*, 216–227. [CrossRef]
31. Kilham, P.; Hartebrodt, C.; Kändler, G. Generating tree-level harvest predictions from forest inventories with random forests. *Forests* **2018**, *10*, 20. [CrossRef]
32. Weiskittel, A.R.; Crookston, N.L.; Radtke, P.J. Linking climate, gross primary productivity, and site index across forests of the western United States. *Can. J. For. Res.* **2011**, *41*, 1710–1721. [CrossRef]
33. Mitsopoulos, I.; Xanthopoulos, G. Effect of stand, topographic, and climatic factors on the fuel complex characteristics of Aleppo (Pinus halepensis Mill.) and Calabrian (Pinus brutia Ten.) pine forests of Greece. *For. Ecol. Manag.* **2016**, *360*, 110–121. [CrossRef]
34. Ye, J.; Wu, B.; Liu, M.; Gao, Y.; Gao, J.; Lei, Y. Estimation of aboveground biomass of vegetation in the desert-oasis ecotone onthe northeastern edge of the Ulan Buh Desert. *Acta Ecol. Sin.* **2018**, *38*, 1216–1225. (In Chinese) [CrossRef]
35. Zhao, X.; Guo, Q.; Su, Y.; Xue, B. Improved progressive TIN densification filtering algorithm for airborne LiDAR data in forested areas. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 79–91. [CrossRef]
36. Pinheiro, J.; Bates, D.; DebRoy, S.; Sarkar, D.; Team, R.C. Linear and nonlinear mixed effects models. *R Package Version* **2007**, *3*, 1–89.
37. Fu, L.; Tang, S. A general formulation of nonlinear mixed effect models and its application. *Sci. Sin. Math* **2020**, *50*, 15–30. [CrossRef]
38. Arnold, T.W. Uninformative parameters and model selection using Akaike's Information Criterion. *J. Wildl. Manag.* **2010**, *74*, 1175–1178. [CrossRef]
39. Fredensborg Hansen, R.M.; Rinne, E.; Skourup, H. Classification of Sea Ice Types in the Arctic by Radar Echoes from SARAL/AltiKa. *Remote Sens.* **2021**, *13*, 3183. [CrossRef]
40. Liaw, A.; Wiener, M.J.R.N. Classification and Regression by random. *Forest* **2002**, *23*, 18–22.
41. Kavzoglu, T.; Colkesen, I. A kernel functions analysis for support vector machines for land cover classification. *Int. J. Appl. Earth Obs. Geoinf.* **2009**, *11*, 352–359. [CrossRef]
42. Shah-Hosseini, R.; Homayouni, S.; Safari, A. A hybrid kernel-based change detection method for remotely sensed data in a similarity space. *Remote Sens.* **2015**, *7*, 12829–12858. [CrossRef]
43. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R Package Version 2015. Available online: https://cran.r-project.org/web//packages/e1071 (accessed on 11 December 2022).
44. Shu, S.; Zhou, X.H.; Shen, X.Y.; Liu, Z.C.; Tang, Q.H.; Li, H.L.; Ke, C.Q.; Li, J. Discrimination of different sea ice types from CryoSat-2 satellite data using an Object-based Random Forest (ORF). *Mar. Geod.* **2020**, *43*, 213–233. [CrossRef]
45. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S-PLUS*; Springer Science & Business Media: Berlin, Germany, 2013.
46. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobotics* **2013**, *7*, 21. [CrossRef]
47. Greenwell, B.; Cunningham, B.B.; Developers, G. gbm: Generalized Boosted Regression Models; R Package Version 2.1.8.1. 2022. Available online: https://CRAN.R-project.org/package=gbm (accessed on 11 December 2022).
48. Kuhn, M. caret: Classification and Regression Training; R Package Version 6.0-93. 2022. Available online: https://CRAN.R-project.org/package=caret (accessed on 11 December 2022).
49. Liang, T.; Yang, S.; Feng, Q.; Liu, B.; Zhang, R.; Huang, X.; Xie, H. Multi-factor modeling of above-ground biomass in alpine grassland: A case study in the Three-River Headwaters Region, China. *Remote Sens. Environ.* **2016**, *186*, 164–172. [CrossRef]
50. Lyu, X.; Li, X.; Gong, J.; Li, S.; Dou, H.; Dang, D.; Xuan, X.; Wang, H. Remote-sensing inversion method for aboveground biomass of typical steppe in Inner Mongolia, China. *Ecol. Indic.* **2021**, *120*, 106883. [CrossRef]
51. Zhang, C.; Lu, D.; Chen, X.; Zhang, Y.; Maisupova, B.; Tao, Y. The spatiotemporal patterns of vegetation coverage and biomass of the temperate deserts in Central Asia and their relationships with climate controls. *Remote Sens. Environ.* **2016**, *175*, 271–281. [CrossRef]
52. Jin, Y.; Yang, X.; Qiu, J.; Li, J.; Gao, T.; Wu, Q.; Zhao, F.; Ma, H.; Yu, H.; Xu, B. Remote sensing-based biomass estimation and its spatio-temporal variations in temperate grassland, Northern China. *Remote Sens.* **2014**, *6*, 1496–1513. [CrossRef]
53. Banerjee, B.P.; Spangenberg, G.; Kant, S. Fusion of spectral and structural information from aerial images for improved biomass estimation. *Remote Sens.* **2020**, *12*, 3164. [CrossRef]
54. Grüner, E.; Astor, T.; Wachendorf, M. Biomass prediction of heterogeneous temperate grasslands using an SfM approach based on UAV imaging. *Agronomy* **2019**, *9*, 54. [CrossRef]
55. Li, A.; Dhakal, S.; Glenn, N.F.; Spaete, L.P.; Shinneman, D.J.; Pilliod, D.S.; Arkle, R.S.; McIlroy, S.K. Lidar aboveground vegetation biomass estimates in shrublands: Prediction, uncertainties and application to coarser scales. *Remote Sens.* **2017**, *9*, 903. [CrossRef]

56. Ramoelo, A.; Cho, M.A.; Mathieu, R.; Madonsela, S.; Van De Kerchove, R.; Kaszta, Z.; Wolff, E. Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and WorldView-2 data. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *43*, 43–54. [CrossRef]
57. Wang, Y.; Wu, G.; Deng, L.; Tang, Z.; Wang, K.; Sun, W.; Shangguan, Z. Prediction of aboveground grassland biomass on the Loess Plateau, China, using a random forest algorithm. *Sci. Rep.* **2017**, *7*, 6940. [CrossRef]
58. Salehnasab, A.; Bayat, M.; Namiranian, M.; Khaleghi, B.; Omid, M.; Masood Awan, H.U.; Al-Ansari, N.; Jaafari, A. Machine learning for the estimation of diameter increment in mixed and uneven-aged forests. *Sustainability* **2022**, *14*, 3386. [CrossRef]
59. Zhao, Q.; Yu, S.; Zhao, F.; Tian, L.; Zhao, Z. Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments. *For. Ecol. Manag.* **2019**, *434*, 224–234. [CrossRef]
60. Mahesh, B. Machine learning algorithms—A review. *Int. J. Sci. Res.* **2020**, *9*, 381–386. [CrossRef]
61. Adamec, Z.; Drápela, K. Comparison of parametric and nonparametric methods for modeling height-diameter relationships. *IForest* **2017**, *10*, 1–8. [CrossRef]
62. Ray, S. A quick review of machine learning algorithms. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 14–16 February 2019; pp. 35–39. [CrossRef]
63. Rudin, C.; Radin, J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harv. Data Sci. Rev.* **2019**, *1*, 2. [CrossRef]
64. Xiangdong, L. Applications of machine learning algorithms in forest growth and yield prediction. *J. Beijing For. Univ.* **2019**, *41*, 23–36. (In Chinese)
65. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]
66. Thessen, A. Adoption of machine learning techniques in ecology and earth science. *One Ecosyst.* **2016**, *1*, e8621. [CrossRef]
67. Fielding, A.H. *Cluster and Classification Techniques for the Biosciences*; Cambridge University Press: London, UK, 2006. [CrossRef]
68. Miao, X.; Heaton, J.S.; Zheng, S.; Charlet, D.A.; Liu, H. Applying tree-based ensemble algorithms to the classification of ecological zones using multi-temporal multi-source remote-sensing data. *Int. J. Remote Sens.* **2012**, *33*, 1823–1849. [CrossRef]
69. Baccarini, L.M.R.; e Silva, V.V.R.; de Menezes, B.R.; Caminhas, W.M. SVM practical industrial application for mechanical faults diagnostic. *Expert Syst. Appl.* **2011**, *38*, 6980–6984. [CrossRef]
70. Safari, A.; Sohrabi, H.; Powell, S.L. Comparison of satellite-based estimates of aboveground biomass in coppice oak forests using parametric, semiparametric, and nonparametric modeling methods. *J. Appl. Remote Sens.* **2018**, *12*, 046026. [CrossRef]
71. Zeng, N.; Ren, X.; He, H.; Zhang, L.; Li, P.; Niu, Z. Estimating the grassland aboveground biomass in the Three-River Headwater Region of China using machine learning and Bayesian model averaging. *Environ. Res. Lett.* **2021**, *16*, 114020. [CrossRef]
72. Cao, L.; Pan, J.; Li, R.; Li, J.; Li, Z. Integrating airborne LiDAR and optical data to estimate forest aboveground biomass in arid and semi-arid regions of China. *Remote Sens.* **2018**, *10*, 532. [CrossRef]
73. Książek, W.; Gandor, M.; Pławiak, P. Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. *Comput. Biol. Med.* **2021**, *134*, 104431. [CrossRef] [PubMed]
74. Greaves, H.E.; Vierling, L.A.; Eitel, J.U.; Boelman, N.T.; Magney, T.S.; Prager, C.M.; Griffin, K.L. Estimating aboveground biomass and leaf area of low-stature Arctic shrubs with terrestrial LiDAR. *Remote Sens. Environ.* **2015**, *164*, 26–35. [CrossRef]
75. Olsoy, P.J.; Glenn, N.F.; Clark, P.E.; Derryberry, D.R. Aboveground total and green biomass of dryland shrub derived from terrestrial laser scanning. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 166–173. [CrossRef]
76. Ni-Meister, W.; Lee, S.; Strahler, A.H.; Woodcock, C.E.; Schaaf, C.; Yao, T.; Ranson, K.J.; Sun, G.; Blair, J.B. Assessing general relationships between aboveground biomass and vegetation structure parameters for improved carbon estimate from lidar remote sensing. *J. Geophys. Res. Biogeosci.* **2010**, 115. [CrossRef]
77. Kim, S.; McGaughey, R.J.; Andersen, H.-E.; Schreuder, G. Tree species differentiation using intensity data derived from leaf-on and leaf-off airborne laser scanner data. *Remote Sens. Environ.* **2009**, *113*, 1575–1586. [CrossRef]
78. Zhang, K.; Hu, B. Individual urban tree species classification using very high spatial resolution airborne multi-spectral imagery using longitudinal profiles. *Remote Sens.* **2012**, *4*, 1741–1757. [CrossRef]