A Comparison of Models of Stand Volume in Spruce-Fir Mixed Forest in Northeast China

Jiarong Liu, Jingyuan He, Lei Chai, Xun Zhong, Bo Jia and Xinjie Wang

College of Forestry, Beijing Forestry University, Beijing 100091, China; jiarongl@bjfu.edu.cn (J.L.); hejingyuan@bjfu.edu.cn (J.H.); chailei@bjfu.edu.cn (L.C.); zhongxun@bjfu.edu.cn (X.Z.); jiaobo8410@bjfu.edu.cn (B.J.)

* Correspondence: xinjiew@bjfu.edu.cn

Abstract: Based on a multiple linear regression model, random forest algorithm and generalized additive model, a stand volume model was constructed to provide a theoretical basis for sustainable management. A total of 224 fixed plots in the Jingouling forest farm, Wangqing County, Jilin Province, were used as data sources. Specifically, 157 plots were used as training data, and 77 plots were used as test data. The effects of stand structure variables, topography variables, cutting variables, diversity variables and climate variables on stand volume were analyzed. The random forest algorithm explained 95.51% of the stand volume, and the generalized additive model explained 95.45% of the stand volume. Stand structure variables and topography variables had more influence on the stand volume of spruce-fir than other variables. Among the diversity variables, the evenness index, Shannon index and Simpson index had a relatively greater impact on the stand volume. The cutting times and the intensity of the first cutting had a direct relationship with stand volume. The influence of climate variables on the stand volume was relatively small in the study area.

Keywords: linear regression model; random forest; generalized additive model; spruce-fir; stand volume

1. Introduction

Forest stand volume is one of the key forest structural attributes in estimating and forecasting carbon stock [1,2]. Therefore, the use of scientific methods to predict stand volume provides a reliable basis for forest management and planning, and is very important in giving full play to the multifunctional benefits of forests. Traditional stand volume measurements are mainly achieved through field measurements, which have high labor intensity and take a long time [3]. The stand growth model refers to a mathematical equation or a set of mathematical equations used to describe the relationship between tree growth and stand state [4]. Due to the importance of stand volume, the study of stand volume models is the main object in the modeling of stand measurement variables [5–8]. The stand volume is affected by climate, topographic conditions and other variables [9]. It can reflect the comprehensive impact of environmental variables on the forest, so it is of great significance to measure stand volume [10].

Traditional stand volume models, which include competition, individual tree size, site and species diversity [11–14], lack variables reflecting climate and human activities. In recent years, to reflect the impact of climate change on forests, some authors have tried to establish climate-sensitive models by adding climate variables into traditional empirical models [15,16]. Many studies have shown that different climatic conditions have a sustained impact on forest productivity [17,18]. However, studies have proven that in the short term, the impact speed and degree of human activities on vegetation exceed those of climate variables [19]. Most studies have shown that moderate cutting can accelerate stand growth, improve forest composition, and maintain the structure of mixed forest with multiple layers and different ages [20,21].
As for simulating a stand volume model, classical multiple linear regression (MLR) has often failed to explain the discrete component of the data [22]. With the development of computers and statistical modeling techniques, machine learning methods may become more useful for nonlinear relationships and interactions. A machine learning algorithm is an algorithm that automatically obtains rules from data, makes use of the rules to predict unknown data, and can be used for regression or classification problems [23]. Many nonparametric machine learning estimation methods have been introduced into regression statistics to estimate forest parameters [24,25]. The main idea of the random forest (RF) model comes from boosting and bagging algorithms, and the random forest model builds multiple decision trees; that is, through the combination of multiple weak learners, the effect of strong learners can be achieved [26]. At present, RF models are mostly used to estimate biomass by remote sensing, but few studies have been performed to predict stand volume. When the RF algorithm is used to predict stand volume, it is necessary to set the hyperparameter, and the performance of the model obtained is often significantly different with different hyperparameter settings [27]. The generalized additive model (GAM) is a nonparametric extension of the generalized linear model, which uses a smoothing function to replace the regression coefficient to make the model more flexible and uses nonparametric methods to detect the structure of data and determine the structure in the data to obtain better prediction results [28]. The smoothing function of GAM is sufficient to express complex and possibly nonlinear trends when the function form is not clear [29]. In recent years, GAMs have been widely used to study the relationship between climate and vegetation in ecology [30].

**Picea** spp. and *Abies* spp. are the dominant species in a Spruce-fir mixed broad-leaved-conifer forest (hereafter referred to as spruce-fir forest). As one of the main forest types in Changbai Mountain in northeast China, the volume of spruce-fir forests is generally higher than those of other forest types. As an important part of northeast China, the Changbai Mountain forest is one of the regions with the richest forest resources in the same latitude in the world and is also an important production base of wood and forest products in China [31]. Due to the overcutting of natural forest resources for a long time, the existing stands in this area are mostly natural secondary forests with species of *Abies nephrolepis*, *Pinus koraiensis* and other tree species as constructive species after more than ten years of restoration. Jingouling Forest Farm is located on the north slope of Changbai Mountain. In this area, spruce-fir forest is the main forest stand.

Long-term continuous monitoring data of spruce-fir sample plots repeatedly measured for 16 years (1991–2006) in the Wangqing Forest of Jilin Province, northeastern China were used. An MLR model, RF algorithm and GAM were used to build stand volume models with 36 independent variables as stand structure variables, topography variables, cutting variables, diversity variables and climate variables. The results of the three models were compared to obtain the optimal stand volume model, and the relative importance and partial dependence of independent variables affecting stand volume were analyzed.

### 2. Materials and Methods

#### 2.1. Study Area

The study area (130°05’–130°19’ E, 43°17’–43°25’ N) is in Jingouling Forest Farm, Jilin Province, which belongs to the Changbai Mountain system (Figure 1). The landform is low hills with an altitude of 592–784 m. The slope is generally 2° to 16°, with some steep slopes are above 20°. The area has a monsoon climate. During the survey period (1991–2006), the average temperature in July was 20.2 °C, the average maximum temperature was 25.6 °C, and the average minimum temperature was 15.8 °C. Dark brown soils are the dominant soil type throughout the study area. The soil layer in the study area was approximately 40 cm thick.
The vegetation in the study area belongs to the flora of Changbai Mountain, with rich vegetation types and complex structures. The study area is a natural dark coniferous forest dominated by *Picea koraiensis* and *Abies nephrolepis*. Other tree species include *Pinus koraiensis*, *Betula costata*, *Tilia amurensis*, *Ulmus pumila*, and *Acer mono*. The shrub species are *Acer tegmentosum*, *Acer ukurunduense*, *Lonicer japonica*, *Spiraea salicifolia*, and *Corylus heterophylla*. Herbs include *Brachybotrys paridiformis*, *Paris verticillate*, *Carex siderosticta*.

### 2.2. Dataset

Since 1991, 224 fixed rectangular sample plots (20 × 20 m) have been measured, with monitored variables including the aspect, slope and altitude of the sample plot, every 3 years (Figure 1). In each survey, tree species, diameter at breast height (d) and height of each tree with d bigger than 5 cm were recorded. In this study, six measurements of observation data from 1991 to 2006 were selected for modeling. Old trees, diseased trees, separated or naturally broken residual trees in some plots were cut down 1–2 times. The climate data were taken from the monthly data set of China’s surface climate standard values (1991–2006) provided by the National Meteorological Science Data Center (http://data.cma.cn, accessed on 4 March 2021). The Kriging spatial interpolation method in ArcGIS software was used to spatially interpolate the data of Wangqing, Jilin and three surrounding climate stations, Yanji, Dunhua and Suifenhe, to obtain the climate data of plots. The data presented in this study are openly available in FigShare at 10.6084, reference number 20078726.

### 2.3. Variable Selection and Calculation

The stand volume was calculated according to a one-way volume table (Table A1, Appendix A) [32] in the study area. Stand volume is affected by many variables. The selection of variables should not only consider the main variables affecting stand volume but also be easy to obtain. This study mainly introduces stand structure variables, topography variables, cutting variables, diversity variables and climate variables. Topography variables, cutting variables and climate variables have a certain impact on stand structure variables and diversity variables. Cutting promotes the increase of diameter at breast height and height of the stand. However, some studies have shown that when the stand composition is a relatively stable community, cutting does not significantly change the tree species composition of the stand and does not have a significant impact on species diversity [33]. The statistics of each variable are shown in Table 1.
Table 1. Statistical table of each variable.

<table>
<thead>
<tr>
<th>Factor Groups</th>
<th>Variables</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stand structure</td>
<td>D</td>
<td>18.75</td>
<td>3.5</td>
<td>34.94</td>
<td>7.09</td>
<td>Mean quadratic diameter stand (cm)</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>900</td>
<td>322</td>
<td>1950</td>
<td>175</td>
<td>Number of trees per hectare (plant/ha)</td>
</tr>
<tr>
<td></td>
<td>VB</td>
<td>0.25</td>
<td>0.22</td>
<td>1</td>
<td>0</td>
<td>Percentage of broad-leaved tree stock</td>
</tr>
<tr>
<td></td>
<td>SDI</td>
<td>2341.03</td>
<td>630.09</td>
<td>4035.16</td>
<td>475.18</td>
<td>Stand Density Index (trees/ha)</td>
</tr>
<tr>
<td>Topography</td>
<td>ALTITUDE</td>
<td>6.51</td>
<td>0.07</td>
<td>6.66</td>
<td>6.39</td>
<td>Ln (ALTITUDE) (m)</td>
</tr>
<tr>
<td></td>
<td>ASPECT</td>
<td>0.57</td>
<td>0.48</td>
<td>1</td>
<td>−0.95</td>
<td>Sin (ASPECT) (')</td>
</tr>
<tr>
<td></td>
<td>SLOPE</td>
<td>0.17</td>
<td>0.07</td>
<td>0.39</td>
<td>0.04</td>
<td>Tan(SLOPE) ('')</td>
</tr>
<tr>
<td>Cutting</td>
<td>FCI</td>
<td>11.6</td>
<td>11.57</td>
<td>49.55</td>
<td>0</td>
<td>Intensity of first cutting (%)</td>
</tr>
<tr>
<td></td>
<td>SCI</td>
<td>1.36</td>
<td>6.5</td>
<td>49.35</td>
<td>0</td>
<td>Intensity of second cutting (%)</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>0.69</td>
<td>0.56</td>
<td>2</td>
<td>0</td>
<td>Cutting times</td>
</tr>
<tr>
<td></td>
<td>FCT</td>
<td>0.37</td>
<td>0.46</td>
<td>1</td>
<td>0</td>
<td>Time from the first cutting (years)</td>
</tr>
<tr>
<td></td>
<td>SCT</td>
<td>0.02</td>
<td>0.12</td>
<td>1</td>
<td>0</td>
<td>Time from the second cutting (years)</td>
</tr>
<tr>
<td>Diversity</td>
<td>Richness</td>
<td>7.17</td>
<td>1.77</td>
<td>11</td>
<td>2</td>
<td>Species richness index</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>8.05</td>
<td>2.74</td>
<td>20.5</td>
<td>2</td>
<td>Chao1 Index</td>
</tr>
<tr>
<td></td>
<td>ACE</td>
<td>8.97</td>
<td>3.34</td>
<td>29.51</td>
<td>2</td>
<td>ACE Index</td>
</tr>
<tr>
<td></td>
<td>Shannon</td>
<td>1.6</td>
<td>0.3</td>
<td>2.22</td>
<td>0.45</td>
<td>Shannon Index</td>
</tr>
<tr>
<td></td>
<td>Simpson</td>
<td>0.74</td>
<td>0.1</td>
<td>0.88</td>
<td>0.28</td>
<td>Simpson Index</td>
</tr>
<tr>
<td></td>
<td>Pielou</td>
<td>0.83</td>
<td>0.09</td>
<td>1</td>
<td>0.47</td>
<td>Pielou Evenness index</td>
</tr>
<tr>
<td></td>
<td>Equitability</td>
<td>0.6</td>
<td>0.14</td>
<td>1</td>
<td>0.24</td>
<td>Equitability Evenness index</td>
</tr>
<tr>
<td>Climate</td>
<td>minPRES</td>
<td>951.58</td>
<td>3.2</td>
<td>957.24</td>
<td>947.75</td>
<td>Minimum pressure (Pa)</td>
</tr>
<tr>
<td></td>
<td>minTEM</td>
<td>10.63</td>
<td>1.82</td>
<td>14.1</td>
<td>8.9</td>
<td>Minimum temperature (°C)</td>
</tr>
<tr>
<td></td>
<td>maxPRES</td>
<td>971.41</td>
<td>0.92</td>
<td>972.55</td>
<td>970.25</td>
<td>Maximum pressure (Pa)</td>
</tr>
<tr>
<td></td>
<td>maxTEM</td>
<td>32.61</td>
<td>2.96</td>
<td>36.46</td>
<td>28.81</td>
<td>Maximum temperature (°C)</td>
</tr>
<tr>
<td></td>
<td>PREC20_20</td>
<td>132.41</td>
<td>80.75</td>
<td>962.15</td>
<td>204.52</td>
<td>Precipitation at 20–20 h (mm)</td>
</tr>
<tr>
<td></td>
<td>meanPRES</td>
<td>959.99</td>
<td>46.18</td>
<td>964.15</td>
<td>204.52</td>
<td>Mean pressure (Pa)</td>
</tr>
<tr>
<td></td>
<td>mean2WindV</td>
<td>1.88</td>
<td>0.23</td>
<td>2.3</td>
<td>1.61</td>
<td>Two-minute average wind speed (m/120 s)</td>
</tr>
<tr>
<td></td>
<td>meanTEM</td>
<td>20.54</td>
<td>1.59</td>
<td>22.78</td>
<td>18.18</td>
<td>Mean temperature (°C)</td>
</tr>
<tr>
<td></td>
<td>meanVolPRES</td>
<td>19.2</td>
<td>1.64</td>
<td>21.65</td>
<td>16.84</td>
<td>Mean vapor pressure (Pa)</td>
</tr>
<tr>
<td></td>
<td>meanRelHUM</td>
<td>78.79</td>
<td>4.83</td>
<td>85.2</td>
<td>70.28</td>
<td>Average relative humidity</td>
</tr>
<tr>
<td></td>
<td>meanminTEM</td>
<td>15.82</td>
<td>1.14</td>
<td>17.82</td>
<td>14.3</td>
<td>Mean minimum temperature (°C)</td>
</tr>
<tr>
<td></td>
<td>meanmaxTEM</td>
<td>26.31</td>
<td>2.4</td>
<td>29.86</td>
<td>22.87</td>
<td>Mean maximum temperature (°C)</td>
</tr>
<tr>
<td></td>
<td>DzPREFC_0.1</td>
<td>15.47</td>
<td>3.39</td>
<td>20.5</td>
<td>17.75</td>
<td>Daily precipitation (&gt;0.1 mm)</td>
</tr>
<tr>
<td></td>
<td>MonSun</td>
<td>41.87</td>
<td>11.51</td>
<td>56.5</td>
<td>24.75</td>
<td>Percentage of monthly sunshine (%)</td>
</tr>
<tr>
<td></td>
<td>SunH</td>
<td>194.61</td>
<td>53.37</td>
<td>262.5</td>
<td>115.57</td>
<td>Hours of sunshine (H)</td>
</tr>
<tr>
<td></td>
<td>maxWindV</td>
<td>10.65</td>
<td>1.86</td>
<td>14.52</td>
<td>8.5</td>
<td>Maximum wind speed (m/s)</td>
</tr>
<tr>
<td></td>
<td>macPRES</td>
<td>28.91</td>
<td>13.21</td>
<td>44.53</td>
<td>8.34</td>
<td>Maximum daily precipitation (mm/D)</td>
</tr>
</tbody>
</table>

2.3.1. Stand Structure Variables

According to the data of every tree in the plot, the main variables affecting the stand volume included the mean quadratic diameter (\(D\)), the number of trees per hectare (N), the volume ratio of broad-leaved trees (VB) and the stand density index (SDI). \(D\) was calculated by Formula (1), N by Formula (2), and VB by Formula (3). The SDI refers to the number of trees per unit area of stand in standard mean diameter, which was used to assess the degree of interstand crowding. The degree of crowding among trees in stands depends on the number of trees per unit area, the average size of trees and the distribution of trees on the forestland [34]. The SDI was calculated by Formula (4).

\[
D = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{N}}
\]  
(1)

where \(d_i\) is the diameter of the ith tree, and \(N_i\) is the total number of trees in the plot.

\[
N = N_i / A
\]  
(2)

where A is the area of the survey plot.

\[
VB = \frac{V_b}{V}
\]  
(3)

where \(V_b\) is the broad-leaved tree volume in plot, \(V\) is the total volume of the plot.

\[
SDI = N \left( \frac{D}{D_0} \right)^\beta
\]  
(4)
where \( D_0 \) is the standard average diameter, and \( \beta \) is the natural thinning slope of the stand. In China, \( D_0 \) is 20 cm, and \( \beta \) is 1.605 [35].

2.3.2. Topography Variables

The topography variables include slope, aspect, and altitude. Altitude has a significant positive correlation with soil related indicators, so altitude can replace the many soil variables. In addition, considering the interaction among the variables, the combination or transformation form of altitude, aspect and slope is often used [36]. The topography variables selected in this study included \( \ln(\text{ALTITUDE}) \), \( \sin(\text{ASPECT}) \) and \( \tan(\text{SLOPE}) \).

2.3.3. Cutting Variables

The cutting variables mainly considered the first cutting intensity (FCI), the second cutting intensity (SCI), the cutting times (CT), the years after the first cutting (FCT), and the years after the second cutting (SCT).

\[
FCT = 1 / \left(1 + t_1^2\right) 
\]

\[
SCT = 1 / \left(1 + t_2^2\right) 
\]

where \( t_1 \) represents the time from the first cutting and \( t_2 \) represents the time from the second cutting.

2.3.4. Diversity Variables

The species richness index (S), Chao1 index (Chao1), ACE index (ACE), Shannon index (Shannon), Simpson index (Simpson), Pielou evenness index (Pielou) and Equitability evenness index (Equitability) of the \( \alpha \) diversity measure method were used to analyze the diversity variables affecting stand volume. \( S \) is a quantitative index to describe the species richness in a community, in which the simplest way is the number of species in a plot [37]. Chao1 was first proposed by Chao [38] to estimate the total number of species according to two rare species indices. The larger the value of Chao1, the more species are represented. The ACE is a comprehensive index of rare and common species [39]. It is classified as a rare or common species with a number of individuals less than or equal to 10 or greater than 10. The Shannon index takes into account the richness and evenness of the community, which can be expressed as the probability of two individuals being selected randomly in the community [40]. The Simpson index is often used to quantify the biodiversity of a region [41]. Pielou defined evenness as the ratio of measured diversity to maximum diversity [42].

\[
S = N_0 
\]

where \( N_0 \) is the total number of tree species in the plot.

\[
\text{Chao1} = S + \left( S_1^{1/2} / S_2 + 1 \right) 
\]

\[
\text{ACE} = S_2 + S_3 / \left( 1 - n_1 / n_r \right) + F_1 \cdot \gamma_{\text{ACE}} / \left( 1 - n_1 / n_r \right) 
\]

\[
\gamma_{\text{ACE}} = \text{MAX} \left[ S_r \cdot \left( 1 - 0.5 \cdot S_1 + 2.1 \cdot S_2 + \cdots + 10 \cdot 9 \cdot S_{10} \right) \right] 
\]

where \( S_k \) is the number of species with only \( K \) individuals, \( S_a \) is the number of common species, \( S_r \) is the number of rare species, \( n_r \) is the number of rare species, \( F_1 \) is the number of species with only one individual.

\[
\text{Simpson} = p_1^2 + p_2^2 + \cdots + p_i^2 \left( 0 < i \leq S \right) 
\]

\[
\text{Shannon} = -\sum_{i=1}^{S} p_i \log_{e} p_i \left( 0 < i \leq S \right) 
\]
where \( P_i \) is the ratio of the number of individuals of species \( i \) to the total number of individuals in the community, \( P_i = \frac{N_i}{N} \).

2.3.5. Climate Variables

The minimum pressure (minPRES), minimum temperature (minTEM), maximum pressure (maxPRES), maximum temperature (maxTEM), precipitation at 20–20 h (PREC20_20), mean pressure (meanPRES), two-minute average wind speed (mean2WindV), mean temperature (meanTEM), two-minute average wind speed (mean2WindV), mean temperature (meanTEM), mean vapor pressure (meanVaPRES), average relative humidity (meanReHUM), mean minimum temperature (meanminTEM), mean maximum temperature (meanmaxTEM), daily precipitation \( \geq 0.1 \) mm (DaPREC_0.1), percentage of monthly sunshine (MonSun), hours of sunshine (SunH), maximum wind speed (maxWindV) and maximum daily precipitation (macPREC) were taken into account in climate modeling.

2.4. Model Building

2.4.1. Multiple Linear Regression

MLR is used to analyze the linear relationship between independent variables and dependent variables [43]. The MLR can be expressed as:

\[
y = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n + b.
\]

Where \( y \) represents the stand volume in the study area, \( x_1, x_2, x_3, \cdots, x_n \) is the D, N, SDI, etc., \( a_1, a_2, a_3, \cdots, a_n \) is the regression coefficient of different variables, and \( b \) is the equation intercept. When one factor affects the influence of another factor on the dependent variable, we need to add an interaction term to the model to show that there is interaction between the two variables. When constructing the regression model, the correlation between their variables is tested by the variance inflation factor (VIF). The larger the VIF, the more serious the multicollinearity between the independent variable and other independent variables [44]. When the VIF of the variable is greater than 10, the model should be re-modelled [45]. The model residual should obey the normal distribution and the size of residual does not change with the change of variables value level; that is, the homogeneity of variance.

2.4.2. Random Forest Algorithm

RF is an integrated learning method based on a decision tree. This model constructs a series of base learners by resampling and randomly selects \( K (1 \leq k \leq p) \) from all variables at each tree node in the process of constructing these base learners to look for an optimal partition variable [46]. In the process of constructing the RF model, two important hyperparameters need to be set: the number of decision trees (ntree) and the number of random variables (mtry) from tree nodes. Ntree is the number of resampling times. Generally, when ntree is greater than 500, the overall error rate tends to be stable, but it still depends on the specific data [8]. Mtry is the number of variables randomly selected from the total number of independent variables each time the optimal separation effect is sought. For regression problems, the default value of mtry is set to \( 1/3 \) of the total number of parameters (rounded) [28]. Because of the difference in the specific data, the optimal model cannot be obtained by taking the mtry default value, and the mtry still needs to be tuned.

Bootstrapping is used to randomly select ntree self-help sample sets from the original data set and to construct a ntree decision tree. The out-of-bag (OOB) data of ntree are composed of the unselected samples in each sampling as the test samples of the random forest model [47]. OOB data do not participate in the fitting of the training set model, so they can be used as a test sample. The random forest model can quantify the relative importance of independent variables to dependent variables and can obtain the partial dependence graph of dependent variables with independent variables [48,49]. The order of
importance of the influencing variables under RF regression was obtained by the IncMSE method. IncMSE is the average of accuracy reduction, and the IncMSE method follows the principle of controlling variables, randomly assigns the values of each variable several times, makes predictions and fits the original model, and then calculates the mean square error (MSE) between the fitting result and the observation result. The larger the MSE, the more important the influencing factor.

Data analysis was performed in statistical software R3.4.4, in which the stochastic forest model was regressed by calling the “randomForest” software package.

2.4.3. Generalized Additive Model

GAM is a nonparametric extension of the traditional generalized linear model, which can effectively deal with the complex nonlinear relations between explanatory variables and effect variables [50]. GAM can be expressed as $Y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$. $\beta_0$ is the intercept, $f_j$ is the smooth function corresponding to the factor, and $\epsilon_i$ is the random error. Each variable can be calculated by any form of function and finally added up to predict $Y$. The GAM is a model based on nonparametric regression and smoothing techniques, in which the difference between expectation and observation is minimized by means of least squares [49]. It is also required that the prediction variables fitted by the spline function should be smooth at the junction of the nodes [51]. A spline is a piecewise function constructed by a polynomial and has a continuous derivative at the piecewise node to make the node highly smooth. In the GAM, the nonlinear function may make the prediction of $Y$ more accurate, but because of the additive hypothesis, the important interaction may be missing, which can be made up by manually adding interaction items. The local dependency graph was highly interpretable and easy to understand. The partial dependence graph of the dependent variable on one independent variable was calculated without ignoring the effect of other variables on the dependent variable but by considering the average effect of other variables on the dependent variable. Bayesian Information Criterion (BIC) was used for model selection, as proposed by Schwarz in 1978 [52]. When training the model, increasing the number of parameters, that is, increasing the complexity of the model, increases the likelihood function, but it also leads to overfitting. To solve this problem, BIC considers the number of samples. When the number of samples is too large, it can effectively prevent the model complexity caused by too high model accuracy.

Data analysis was performed in statistical software R3.4.4, in which the GAM model was regressed by calling the “gam” software package.

2.5. Model Evaluation

After randomly selecting 70% of the data for training and 30% for testing, the adjustment determination coefficient ($R^2_{adj}$), MSE, root mean square error (RMSE) and mean absolute error (MAE) were used as the basic evaluation indexes. $R^2_{adj}$ is the fitting degree of the fitting value to the observed value after regression. The range of the fitting value is [0,1], and the closer $R^2_{adj}$ is to 1, the better the fitting effect of the regression model. MSE is the average of the sum of squares of residuals, which is one of the most commonly used loss functions in regression. The smaller the RMSE and MAE, the higher the model accuracy. All statistical analyses were performed using R software.

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

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

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|
\]

\[
R^2_{adj} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \left( \frac{n - 1}{n - k} \right)
\]
where $\tilde{y}_i$ is the estimated stand volume data, $\bar{y}$ is the average value of $y_i$, $n$ is the number of samples, and $k$ is the number of variables.

3. Results

3.1. Multiple Linear Regression

The result of the final model was $V = 13.08ALITUDE + 2.34SLOPE - 5.4VB + 0.005SCI - 0.15CT + 0.21SCT + 0.72ASPECT$. The $R^2_{adj}$, RMSE, MSE and MAE of the model were 18.23%, 2.3921, 5.7221 m$^3$ and 1.9383 m$^3$, respectively. The diagnosis diagram of the model is shown in Figure 2a, in which the residual met the normality and homogeneity of variance. The weighted graph of the linear regression model for predicting the stand volume of the spruce-fir forest is based on the influencing variables. In Figure 2b, −5 and 15 represent weight estimate, and the line segment represents the drop within the 95% confidence interval; the longer the line, the wider the influence range of stand volume.

3.2. Random Forest

As shown in Figure 3, the overall error of ntree tended to be stable after 3000, so the ntree value was 3000. The OOB error was 0.0967 when mtry was 18, the OOB error was 0.0404 when mtry was 27, and the OOB error was 0.0367 when mtry was 36, so the mtry was 27. All the influencing variables were input into the RF model, $R^2_{adj}$, RMSE, MSE and MAE of the RF model were 95.51%, 0.6111, 0.3735 m$^3$ and 0.4744 m$^3$, respectively. The results of the model fitting are shown in Figure 4.

![Figure 2](image_url)  
(a) Multiple linear regression. Diagnosis diagram of Fit1. (b) Multiple linear regression. The weighted graph of Fit1.

The influence of each variable on stand volume can be directly seen through the weight graph. Among the seven variables selected by the MLR model, topography variables included ASPECT, ALTITUDE, and SLOPE; stand structure variables included VB; cutting variables included CT, SCT, and SCI; and no diversity variables and climate variables. With increasing altitude, the stand volume increased. In addition, ASPECT and SLOPE also had important influences on stand volume. There was a negative correlation between VB and stand volume. The higher the proportion of broad-leaved trees in the stand, the lower the total stand volume. Cutting variables had little effect on the volume.

3.2. Random Forest

As shown in Figure 3, the overall error of ntree tended to be stable after 3000, so the ntree value was 3000. The OOB error was 0.0967 when mtry was 18, the OOB error was 0.0404 when mtry was 27, and the OOB error was 0.0367 when mtry was 36, so the mtry was 27. All the influencing variables were input into the RF model, $R^2_{adj}$, RMSE, MSE and MAE of the RF model were 95.51%, 0.6111, 0.3735 m$^3$ and 0.4744 m$^3$, respectively. The results of the model fitting are shown in Figure 4.
The IncMSE ranking results are shown in Figure 5. SDI and $\bar{D}$, which ranked first and second, had importance values of 420.6307% and 204.1949%, respectively. However, the importance value of N ranked third at only 42.3560%, which fully indicated that SDI and second, had importance values of 420.6307% and 204.1949%, respectively. However, the influencing variables in the bottom 10 were SCI, maxTEM, meanReHUM, DaPREC, mean2WindV, minTEM, SunH, SCT, maxWindV and Monsun. These variables were cutting and climate variables, which indicate that...
the interpretation ability of local-scale climate variables to stand volume was limited. The effects of SCI and SCT on stand structure volume were less than those of FCI and FCT. The relative importance of the four stand structure variables to the stand volume was as high as 69.64%, including SDI420.63%, D204.19%, N44.34%, and VB27.27%. The relative importance of the topography variables to the stand volume was 70.28%, of which ASPECT was 35.46%. The relative importance of the five cutting variables to the stand volume was 37.73%, and the relative importance of 17 climatic variables to the stand volume was 101.31%.

Figure 5. Random Forest. Scatter plot of the importance measure of the output variable.

3.3. Generalized Additive Model

The number of variables predicted by the optimal model under the BIC criterion was 26 (Figure 6), including four stand structure variables, D, N, VB, SDI, three topography variables, ALTITUDE, ASPECT, SLOPE, three cutting variables, CT, FCT, SCT, five diversity variables, Richness, Chao1, Shannon, Simpson, Pielou, eleven climate variables, minPRES, minTEM, maxPRES, maxTEM, meanVaPRES, meanReHUM, meanminTEM, DaPREC_0.1, SunH, maxWindV, maxPREC. D, SDI, Richness, Shannon, Simpson, ALTITUDE, ASPECT, SLOPE, and maxPRES were used to construct the smooth spline function. The other variables were treated as linear relations, and the GAM was built again. $R^2_{\text{adj}}$, RMSE, MSE, and MAE were 95.45%, 0.5453, 0.2974 m³ and 0.4213 m³, respectively. Therefore, the optimal GAM is $V = f(s(D), N, VB, s(SDI), D, C, FCT, SCT, s(Richness), Chao1, s(Shannon), s(Simpson), Pielou, s(ALTITUDE), s(ASPECT), s(SLOPE), s(minPRES), s(minTEM), s(maxPRES), s(maxTEM), s(meanVaPRES), s(meanReHUM), s(meanminTEM), DaPREC_0.1, SunH, s(maxWind), s(maxPREC))$. 

Figure 6. Generalized Additive Model. Number of prediction variables of optimal model under the BIC criterion. The red dot is the best number of variables.
D was less than 17 cm, the stand volume decreased with the increase of D and SDI but decreased with increasing VB. The diversity D, VB and SDI were more important than VB in the RF model and GAM. In addition, Among the diversity variables, Chao1 and Shannon had a positive correlation with V, while Pielou, Richness and Simpson had a negative correlation between SCT and V. Cutting was beneficial to the increase of stand volume. Among the diversity variables, Chao1 and Shannon had a positive correlation with V, while Pielou, Richness and Simpson had a negative correlation with V. Among the 11 climatic variables, only maxPRES had a nonlinear relationship with V while the other variables had a linear relationship, minPRES, minTEM, meanReHUM, meanminTEM, SunH, maxPREC had a negative correlation with V, and maxTEM, meanVaPRES, DaPREC_0.1, maxWindV and V were positively correlated.

3.4. Comparison and Analysis of Model Results

The fitting precision of the RF model and the GAM was obviously higher than that of the MLR model, and the seven variables in the MLR were insufficient to explain the stand volume. The $R^2_{adj}$ of the RF model was 95.51%, and the $R^2_{adj}$ of the GAM was 95.45%. However, the RF model included all variables, and the GAM only used 26 variables to explain the volume. Therefore, it is optimal to use GAM to predict stand volume. The topography variables, ALTITUDE, ASPECT and SLOPE, were of high importance in the three models, indicating that topography variables had important effects on the spruce-fir stand volume. Both the MLR and RF models showed that local climate variables had limited influence on the stand volume, while cutting variables had more influence on the stand volume than diversity variables. Among the 11 climate variables in the GAM, only maxPRES had a nonlinear relationship with V, and maxPRES was the 13th most important in the RF model. The results showed that the maxPRES of climatic variables had an important effect on the spruce fir stand volume. Several variables with high importance rankings in the RF model were involved in both the MLR model and GAM. However, the response of the same variable to the dependent variable was different in the three models. For example, only VB was considered in the MLR model of stand structure variables, but N was more important than VB in the RF model and GAM. In addition, $D$, VB and SDI were the three important stand structure variables, but it should be noted that the stand volume increased with increasing $D$ and SDI but decreased with increasing VB. The diversity variables varied greatly among the fitting results of the three models. The most important diversity factor in the RF model was Equitability, and four of the top ten variables in the RF were diversity variables, while only FCI ranked higher in the cutting variables. The results

![Figure 6](image_url)
show that the effect of diversity variables on stand volume was higher than that of cutting variables in the RF model.

![Figure 7](image_url)

Figure 7. Generalized Additive Model. The local dependence diagram of each influencing factor. The blue line represents the correlation between independent variables and dependent variables, the dotted line indicates the 95% confidence interval.

4. Discussion

4.1. Model Discussion

MLR has a fast modeling speed and does not need very complex calculations. However, the explanatory ability of the model is better when the independent variable has a significant impact on the dependent variable and has a close linear correlation. When using RF and MLR to predict Eucalyptus stand volume and other related tree structure attributes, the \( R^2 \) of RF was significantly higher than that of MLR [53]. There is not a significant linear correlation between stand volume and stand structure, topography and other variables. The performance of RF is better and has great advantages over other algorithms, but we could not obtain the final model results. The GAM overcomes the shortcomings of MLR. Huang used the GAM to well explain the impact of climate change on the stand volume of *Pinus taeda* in the northern United States [54].

4.2. Stand Structure Variables

The results showed that the stand volume was greatly influenced by stand structure variables. This is consistent with some existing research results [55,56], which reflected...
the change of stand volume was determined by stand biometric parameters. When the trees are not affected by competition, the greater the SDI, the greater the stand volume. In contrast, if there are too many trees per unit area, the competition among individual trees is more intense, which leads to a decrease in stand volume. The effect of D on stand volume is closely related to N and tree species composition. When the D of the stand is small, the forest may be in the early stage of succession. At this time, the number of trees per unit area is larger, and the ratio of broad-leaved trees is likely to be relatively higher. Under the condition of D less than 17 cm, the competition pressure of tree species in the forest leads to a decrease of stand volume, and then with the increase of D, the stand structure tends to be stable, and the stand volume increases.

4.3. Topography Variables

Forest trees are influenced not only by the genetic variables of trees but also by the environmental conditions. Topography is an important factor that determines the development of trees, including altitude, slope and aspect. The slope direction affects the stand volume indirectly through the influence of light, and southern slopes have more light than the northern slope. Therefore, the volume of the spruce-fir forest on the southwestern slope was higher than that on the northeastern slope. With increasing altitude, the stand volume also increased. This is different from Kucharzyk's study [57], which showed that a significant decline in volume growth was found at higher altitudes. The reason may be that the altitude span of the study area is small. Further analysis showed that the proportion of broad-leaved trees increased with decreasing altitude. Therefore, it can be inferred that altitude affects stand volume by affecting tree species composition. With increase in the slope, the stand volume decreased as a whole, but there were some inflection points in the middle. It has been found that the effect of the slope aspect on the relative growth rate of the D of large trees is greater than that of young trees [58]. Therefore, the inflection point may be due to the different stages of stand development. To sum up, there are great differences in topographic conditions in different study areas. In order to adapt to different habitats, plants also have different growth mechanisms. It can be seen that the interaction between stand volume and various variables is very complex.

4.4. Cutting Variables

Cutting had a direct effect on the stand volume. The main variables influencing the stand volume were the FCI and the CT. There was a positive correlation between CT and stand volume. This is contrary to the research results of Gupta [59], which concluded that fifteen years post-thinning, stand volume was less in the thinned stands relative to the unthinned controls. The trees retained after cutting were better, which further promoted the growth of the stand, but this took some time. The effect of the first cutting was better than that of the second cutting. With the increase in the FCT, the stand volume also increased. However, there was a negative correlation between SCT and stand volume, indicating that the volume of the stand did not exceed the state before cutting. The time after cutting also had some influence on the stand volume. With increasing time, the effect of SCT was the same as that of FCT. Cutting affected the speed of the change in volume of small and medium diameter class to large and super large diameter class, but the degree of influence was not large.

4.5. Diversity Variables

The higher the diversity index, the richer the species in the community, and the better the stand structure [60]. The importance values of Equitability and Pielou were higher than those of Simpson, Shannon and ACE; Shannon were higher than those of Chao1 and S. The evenness index took into account both the diversity index and the richness index, so the importance value became higher. In GAM, there was a positive correlation between Chao1, Shannon and V. Many studies have shown that high species diversity was not always accompanied by improved stand volume and carbon stock [61,62].
4.6. Climate Variables

The influence of climate variables on the stand volume was not significant at a small scale, which may be due to the plots being distributed within a small area with low climate variation. However, the annual temperature and precipitation had a relatively significant influence on the stand volume [5]. In this study, climate variables mainly included pressure, temperature and precipitation. Among them, both maxPRES and minPRES had a negative correlation with the stand volume, and the uneven distribution of pressure would have a certain impact on the trees; the higher the meanVaPRES, the higher the temperature and the better the growth of the spruce-fir forest, which was consistent with the effect of maxTEM on the stand volume. This is consistent with the results of Magruder [63]. DaPREC_0.1 was positively correlated with stand volume, while maxprec was negatively correlated with stand volume. This is due to the increase of precipitation can improve soil humidity and atmospheric humidity, promote the production of new leaves and branches so as to promote the growth of trees [64,65]. However, aerobic respiration of tree roots is affected if the precipitation is sufficient or too much; however, at this time, precipitation was not correlated with radial growth of trees or even negatively correlated, especially in humid regions [66].

5. Conclusions

In this study, the data of six measurements from 1997 to 2006 of a spruce-fir mixed forests in northeast China were used as research materials for modeling by MLR, RF model and GAM. The fitting precision of the three models was compared. The results showed that GAM had higher fitting accuracy and adaptability and better prediction ability for stand volume. The effects of stand, topography, cutting, diversity and climate variables on the stand volume were analyzed. Stand structure variables and topography variables were the dominant variables affecting the stand volume. The influence of diversity variables, cutting variables and climate variables were relatively limited. The spatial difference in climate data may not be fully reflected due to the small distance between sample plots.

Due to the short research cycle, this paper is only a preliminary research result. Reliable results can only be obtained by long-term observation, especially climate and cutting effects. Next, we can try to use a wider range of sample plot data to explore the impact of climate and cutting on stand volume. In addition, since all plots are established in a relatively small study area, spatial autocorrelation should also be considered.

Author Contributions: Writing original draft preparation, J.L.; methodology, J.H., X.W.; software, J.H.; investigation, L.C., B.J.; writing review & editing, X.Z.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Plan (2017YFC050410101).

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

Appendix A

Table A1. One-Way Volume Table of Wangqing Forest Farm.

<table>
<thead>
<tr>
<th>Tree Species</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conifer</td>
<td>V = 0.0000578396(−0.1349 + 0.9756D)1.8892(46.4026 − 2137.9188/(D + 47))10.98755</td>
</tr>
<tr>
<td>Class I Broad-leaved</td>
<td>V = 0.000053309(−0.1659 + 0.9734D)1.88452(29.445 − 468.9247/(D + 15.7))10.99834</td>
</tr>
<tr>
<td>Class II Broad-leaved</td>
<td>V = 0.000048841(−0.1970 + 0.9728D)1.84048(24.8174 − 402.0877/(D + 16.3))11.05252</td>
</tr>
<tr>
<td>Oak</td>
<td>V = 0.00006125334(−0.1791 + 0.9737D)1.8810091(23.4292 − 353.6657/(D + 15))10.94662565</td>
</tr>
</tbody>
</table>

Conifer: including *Pinus koraiensis*, *Picea koraiensis*, *Pinus sylvestris* var. *mongolica*, *Larix gmelinii*, *Abies nephrolepis*.

2. Class I Broad-leaved: including *Fraxinus mandshurica*, *Juglans mandshurica*, *Phellodendron amurense*, *Tilia amurensis*, *Ulmus pumila*, *Betula costata*, *Betula platyphylla*, *Populus spp*.

3. Class II Broad-leaved: *Acer mono*.

4. Oak: *Quercus spp*.
References

5. Tang, Y.; Shao, Q.; Shi, T.; Wu, G. Developing growth models of stand volume for subtropical forests in karst areas: A case study in the Guizhou Plateau. Forests 2021, 12, 83. [CrossRef]
10. Liu, J.; Feng, Z.; Mannan, A.; Khan, T.U.; Cheng, Z. Comparing non-destructive methods to estimate volume of three tree taxa in Beijing, China. Forests 2019, 10, 92. [CrossRef]
24. Lu, J.; Feng, Z.; Zhi, Y. Estimation of forest biomass and carbon storage in China based on forest resources inventory data. Forests 2019, 10, 650. [CrossRef]


42. Richardson, S. Species diversity and pattern-diversity in the study of ecological succession. *Journal of theoretical biology* 1966, 10, 370–383. [CrossRef]


45. Ishwaran, H.; Malley, J.D. Synthetic learning machines. *BioData Min.* 2014, 7, 1–12. [CrossRef]


51. Ishwaran, H.; Malley, J.D. Synthetic learning machines. *BioData Min.* 2014, 7, 1–12. [CrossRef]


60. Soca, J. Effect of stand biometrical parameters and selected environmental factors on the current annual volume increment of mountain spruce stands. *Sylvan. 2009*, 153, 231–39. [CrossRef]


64. Gupta, S.D.; Pinnen, B.D.; McCready, T. Commercial thinning effects on growth, yield and mortality in natural lodgepole pine stands in Alberta. *For. Chron.* 2020, 96, 111–120. [CrossRef]

65. Park, J.; Kim, H.S.; Jo, H.K.; Jung, I.B. The influence of tree structural and species diversity on temperate forest productivity and stablility in Korea. *Forests* 2019, 10, 1113. [CrossRef]


