forests

MDPI

Article

Site Quality Classification Models of *Cunninghamia Lanceolata* Plantations Using Rough Set and Random Forest West of Zhejiang Province, China

Chen Dong [1,2], Yuling Chen [3,4,*], Xiongwei Lou [1,2,*], Zhiqiang Min [1,2] and Jieyong Bao [5]

1 College of Mathematics and Computer Science, Zhejiang A&F University, Hangzhou 311300, China
2 State Key Laboratory of Forestry Intelligent Monitoring and Information Technology, Zhejiang A&F University, Hangzhou 311300, China
3 College of Environmental and Resources Science, Zhejiang A&F University, Hangzhou 311300, China
4 State Key Laboratory of Subtropical Silviculture, Zhejiang A&F University, Hangzhou 311300, China
5 Hundsun Technologies Inc., Hangzhou 310051, China
* Correspondence: chenyuling92@163.com (Y.C.); lxw@zafu.edu.cn (X.L.); Tel.: +86-131-2137-8389 (Y.C.); +86-131-2137-8389 (X.L.)

Abstract: The site quality evaluation of plantations has consistently been the focus in matching tree species with sites. This paper studied the site quality of Chinese fir (*Cunninghamia lanceolata*) plantations in Lin'an District, Zhejiang Province, China. The site quality model was constructed using the algebraic difference approach (ADA) to classify the site quality grades. The rough set algorithm was used to screen out the key site factors affecting the site rank of Chinese fir plantations. Site quality classification models based on random forest were established, and the importance of key site factors was evaluated. The results are as follows. The random forest model based on the rough set algorithm had small scale and low complexity, and the training and testing accuracies of the model were 92.47% and 78.46%, respectively, which were better than the model without attribute reduction. The most important factors affecting Chinese fir growth in the study area were the slope aspect, slope grade, and canopy closure. The least important factors were the humus layer thickness, soil layer thickness, naturalness, and stand origin. The attribute reduction method proposed in this study overcame the subjectivity of traditional site factor selection, and the site quality classification model constructed improved the model accuracy and reduced the complexity of the algorithm. The methods used in this study can be extended to other tree species to provide a basis for matching tree species with sites and to improve the level of forest management in the future.

Keywords: *Cunninghamia lanceolata* plantations; site quality classification models; site quality evaluation; rough set; random forest

check for updates

Citation: Dong, C.; Chen, Y.; Lou, X.; Min, Z.; Bao, J. Site Quality Classification Models of *Cunninghamia Lanceolata* Plantations Using Rough Set and Random Forest West of Zhejiang Province, China. *Forests* **2022**, *13*, 1312. https://doi.org/10.3390/f13081312

Academic Editor: Joana Amaral Paulo

Received: 12 July 2022
Accepted: 16 August 2022
Published: 17 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Forestation is an important natural and strategic resource in China. The development of forestland resources in Zhejiang Province is relatively high, but its quality has been disregarded in the long-term national economic development process. Therefore, protecting forestland and improving its quality have been important measures to balance rapid economic growth and ecosystem protection. The accurate evaluation of forest site quality is an important guarantee for matching tree species with sites and establishing plantation management measures scientifically. Such an evaluation is a significant premise for realizing the scientific, reasonable, and efficient utilization of forestland. If a scientific and accurate evaluation system of forest site quality can be proposed, then it will have an immense impact on improving the productivity and sustainable development of plantation forests [1].

At present, numerous studies have been conducted in the field of the site quality evaluation of plantation forests. Site quality evaluation aims to classify the suitability or

potential productivity of forestland by collecting relevant data with generalized mathematical methods to take forest management measures according to different classification results [2]. The selection of site factors and site evaluation methods is the key to solving the problem of the site quality evaluation of the plantations.

(1) Site factors

The site quality of plantations can be reflected by some site factors [3]. Given that measuring all of the numerous site factors is impossible, identifying which factors are used as the basis of site quality evaluation has constantly been the research focus in this field. The existing research results have shown that commonly used site factors at present are mainly summarized as topography, soil, climate, and vegetation factors. Traditional forest site quality evaluation involves obtaining relevant site factors through ground measurement and dividing site types, thereafter according to the combination of different site factors. The research on dominant tree species has been conducted particularly in regions with relatively consistent climatic conditions, in which the concern is on the relationship between the local soil or topographic factors and tree height or site index [4–6]. Moreover, some scholars have indicated that ground survey data are mostly discrete non-numerical data, thereby reducing the convergence and stability of the site quality model. Therefore, the focus of site factor acquisition has shifted to climate and plant biological factors. For climate factors, temperature, humidity, precipitation, and dryness are the most selected factors applied to the site quality evaluation of *Pinus koraiensis*, *Picea asperata*, *Fagus longipetiolata*, *Quercus mongolica*, and other tree species [7–10]. For vegetation factors, the height and coverage of understory vegetation are often used to construct a site index model [11].

In collecting site factors, topography, soil, and understory vegetation factors can be obtained through the Forest Management Inventory System of China or field investigation. Studies have shown that soil chemical properties and nutrients such as soil PH, nitrogen, phosphorus, potassium, and other elements directly affect the growth of trees [12]. However, determining these factors is time-consuming and costly. In local stands, the effects of climate change on tree growth are not significant [13]. Therefore, the effects of climate can be markedly important at the landscape and regional scales, while topography and soil can be significant at the local scale. Therefore, in selecting site factors of local stands, choosing factors that can affect the growth of the stand and are easy to determine and measure is better.

At present, selecting site factors in site quality model construction is often subjective. ANKIWAN [14] estimated the average height growth model by using 32 site environmental factors such as topography, gradient, effective soil depth, and the average height of five dominant trees in the Jeju special self-governing province and southern area. Chen et al. [15] selected eight indices, namely, geomorphology, slope aspect, slope position, slope degree, altitude, soil type, soil parent materials, and soil thickness to study the site quality classification rules of Chinese fir and Masson's pine using the decision tree algorithm. Some scholars have used a series of mathematical methods to reduce the dimension of numerous site factors. Guo et al. used principal component analysis (PCA) to select eight main relevant factors (i.e., slope, position, aspect, soil type, humus thickness, soil thickness, landform, and altitude) affecting tree growth from the original 16 site factors, and classified the site quality grade using the comprehensive fuzzy method [16]. Lv et al. selected nine indicators (e.g., soil thickness, soil type, aspect, and position) and reconstructed a stand index model through expert scoring and weighting via the Delphi method [17]. Quichimbo et al. used the CART method to reduce the dimension of subjective soil factors and to analyze the relationship between the dominant height and soil factors [4]. Site factor selection is a multi-attribute fuzzy decision-making problem, and the relationship among factors is constantly complex. Hence, finding key factors affecting stand growth is difficult. The site factor selection method typically relies on prior knowledge, and the results are subjective. To solve this problem, this study used rough set theory to reduce the dimension of site factors. Rough set theory, which was proposed by Professor Pawlak in 1982, is a mathematical tool that can quantitatively deal with inaccurate, inconsistent, and

incomplete information [18]. Rough set can accurately calculate the attribute factors closely related to the decision attribute from the data level without prior knowledge and remove redundant information on the premise of keeping the original classification ability.

(2) Site quality evaluation method

The traditional way to evaluate the stands' site quality mainly includes direct and indirect evaluations [3]. Site class (SC) and site index (SI) methods, as direct evaluation methods, are often used to evaluate the site quality of plantations [19]. That is, the site quality of plantations is evaluated according to the average height and dominant height of the stands. Mathematical methods such as the guide curve model [20], random effect model [21], algebraic differential approach (ADA) [22], generalized ADA (GADA) [23,24], parameterization model [25], and mixed effect model [26] are commonly used to establish the SC and SI models. The indirect method mainly establishes the multiple regression equation between the tree height and site factors to evaluate the growth potential of trees; the quantity theory I model is a typical method and is mostly used in the quality evaluation of non-forested sites [27], where its basic principle is to convert the data of each plot to 0–1 according to the sub-classes of the site factors (i.e., the site factor of slope position is divided into three sub-classes, namely upper, middle, and lower), so as to construct the regression equation of sub-small classes site factors and dominant heights. In addition, the functional relation of the site index between tree species is used to evaluate the site quality, but the accuracy of the results depends on the similarity of the growth types of tree species [28].

Site quality evaluation methods are based on traditional linear or nonlinear modeling methods, which need to have certain statistical assumptions such as data independence, normal distribution, and equal variance. However, the relationship between tree growth and site factors is typically complex and nonlinear, and most forest growth data do not meet this assumption, thereby resulting in difficulty in providing accurate prediction results [29]. For example, biased estimation or invalid prediction would easily occur when traditional regression analysis is used [30]. Site factors screened by PCA can effectively simplify the data structure, but the cumulative contribution rate of the first several principal component factors is consistently low and the key factors cannot be easily determined. The application of quantity theory I can effectively deal with discrete attribute factors, but it depends on the long-term observation data.

In recent years, machine learning, as a new artificial intelligence technology, has gradually entered the field of forestry scientific research to satisfy the needs of forestry production [31]. Compared with traditional statistical models, the machine learning method has no assumptions on the distribution form of data, can considerably process data with high dimensions and complex nonlinear interactions, and can deeply mine valuable information [32]. Furthermore, machine learning models based on recursion, resampling, averaging, and randomization can reveal the hidden structure in the stand data, obtain accurate site quality prediction, and discover new relationships [33]. In machine learning technology, random forest can effectively deal with nonlinearity, interaction, collinearity, and other problems, and can effectively avoid multiple fitting [34]. Moreover, random forest can be used for regression, classification, and prediction, and can also measure the importance of the variables.

Chinese fir is one of the major plant species in Southern China, particularly in Zhejiang Province, and exhibits characteristics such as fast growth, high yield, good material, and significant economic value [35]. Research on trees and stands is essential for Chinese fir planation management in the region. The motivation of the present study is to explore the role of rough set in site factor dimension reduction, develop site quality classification models for Chinese fir using rough set theory and random forest algorithm in Lin'an District, Zhejiang Province, and compare the accuracy of classification models under different site factors. Accordingly, the influence of key site factors on Chinese fir site quality is explored, and the comprehensive evaluation of the forest quality grade is realized.

2. Materials and Methods

2.1. Data

Lin'an District is in the west of Hangzhou City in Zhejiang Province, China from 118°51′ to 119°52′ east longitude and from 29°56′ to 30°23′ north latitude. Data were derived from the dynamic monitoring data of forest resources in Lin'an District from 2009 to 2012, which was the annual dynamic monitoring system of counter-level forest resources established based on the planning and design investigation of forest resources in Zhejiang Province. Investigation factors were based on sub-compartments containing 75 investigation attributes such as the basic information of sub-compartments, site factors, stand factors, management measures, disease, insect, and fire information.

We selected sub-compartment data with the dominant species of Chinese fir as the dataset. In the selection of site factors, we attempted to choose factors that could affect the growth of the forest stand and could be easily obtained. A total of 20 factors were chosen as the initial site factors based on existing research, landform, altitude, slope direction, slope position, slope gradient, soil types, soil texture, soil layer thickness, humus layer thickness, undergrowth vegetation species, undergrowth vegetation height, undergrowth vegetation coverage, plant community structure, naturalness, forest class, forest protection grade, land type, age group, stand origin, and canopy closure. Among the site factors, except altitude, the undergrowth vegetation height, understory vegetation coverage, and canopy density, which belong to continuous data, other site attribute data were classified and assigned based on the planning and design investigation of forest resources in Zhejiang Province.

Given that seedlings were in the recovery and rooting phases, Chinese fir truly entered the fast-growing phase five years later. The related literature has indicated that only forest stands with canopy densities exceeding 0.2 could sufficiently embody the forest tree growth status [36]. Thus, small-class data with ages and canopy densities below 5 years and 0.2, respectively, were excluded in this study. In addition, the data integrity and consistency were checked, and abnormal data were excluded by taking thrice the standard deviation as the criterion. A total of 1903 sub-compartment data were obtained through data processing.

Sub-compartment site data were obtained as depicted in Table 1.

Table 1. The general site information of Chinese fir stands in Lin'an District.

Factor Nos.	Site Factors	Related Values
1	Landform	Medium hills, lowland, irregular hillslopes
2	Altitude (m)	10–1104 m
3	Slope direction	East, south, west, north, northeast, southeast, northwest, southwest
4	Slope position	Ridge, upper, middle, lower, valley, whole
5	Slope gradient	Flat, gentle, inclined, steep, abrupt, dangerous
6	Soil types	Red soil, yellow soil, limestone soil, purplish soil
7	Soil texture	Sandy soil, loamy soil, clay
8	Soil layer thickness	Thick, medium, thin
9	Humus layer thickness	Thick, medium, thin
10	Undergrowth vegetation species	Grass cluster, shrub, bush wood, miscan stem, bamboo fungus
11	Undergrowth vegetation height (cm)	0–85 cm
12	Undergrowth vegetation coverage	0%–90%
13	Plant community structure	Complete structure, relatively complete structure, simple structure
14	Naturalness	Classes I, II, III
15	Forest class	Public welfare forests, commercial forests
16	Forest protection grade	Grades I, II
17	Land type	Highwood land, open forest land
18	Age group	Young forest, middle-aged forest, near mature forest, mature forest
19	Stand origin	Natural forest, plantation
20	Canopy closure	0–0.85

2.2. Methods

In this study, the site index model of Chinese fir was constructed using ADA to determine the site quality grade. A total of 20 site factors were reduced by rough set according to the site quality grade after the discretization of continuous data and balance of the sample data. Classification models based on random forest were likewise carried out with site quality grade as a dependent variable and site factors as independent variables. Unlike in previous studies, there are two ways to select independent variables in this study: all 20 site factors considered as independent variables formed scheme A and the key site factors after rough set were considered as independent variables (scheme B). The two schemes were compared, the best method model was selected for the quality classification of Chinese fir, and the importance of the site factors was comprehensively evaluated. The model building process is shown in Figure 1.

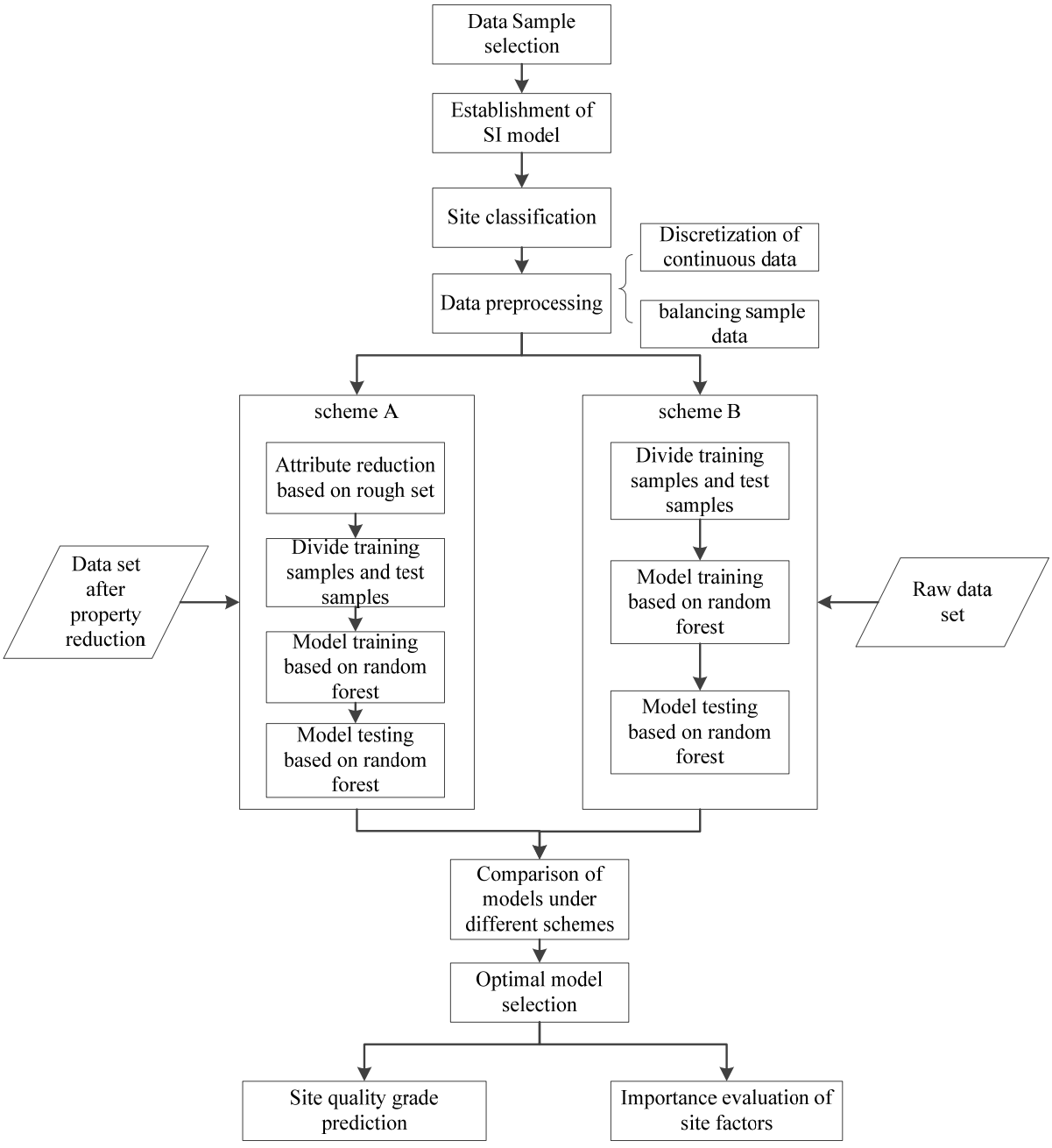


Figure 1. The flowchart of the modeling process.

2.2.1. Site Quality Grade Classification

SC and SI are commonly used in site quality evaluation in China [37]. Compared with SC, SI is widely used because of its specific mathematical expression and less artificial interference in the modeling process [38,39]. Therefore, the current study used ADA to construct the SI model of Chinese fir as a judgment index of the site conditions.

(1) Establishment of the SI model

ADA is one of the common methods used to establish forest stand site indices. Its principle is to select a theoretical growth equation as the basic (fundamental) equation and select one parameter in the equation as the element elimination parameter, thus elimination of the equation to obtain a difference equation including two groups of dependent and independent variables [40].

In this study, the Richards equation was used as the basic equation to establish the difference equation. The equation can be expressed as follows:

$$HT = a(1 - exp(-ct))^b \quad (1)$$

where HT is the dominant height of the sub-compartment; t is the age of the sub-compartment; and a, b, and c are the parameters. In Equation (1), parameter a represents the maximum potential growth of trees, parameter c represents the growth rate of trees, and parameter b is used as an elimination parameter to obtain the elimination. The converted difference equation is as follows:

$$HT_2 = a(HT_1/a)^{\frac{ln(1-exp(-ct_2))}{ln(1-exp(-ct_1))}} \quad (2)$$

where HT_1 and HT_2 are the dominant tree heights of the sub-compartment in years t_1 and t_2, respectively; other variables are as previously defined. The specific construction process of the difference equation can be referred to in [40].

We selected the data of the dominant tree heights and age of the Chinese fir sub-compartment in 2009 and 2012. HT_1 and HT_2 in Equation (2) represent the dominant tree heights of Chinese fir sub-compartments in 2009 and 2012, respectively; t_1 and t_2 represent the average age of Chinese fir sub-compartments in 2009 and 2012, respectively. SPSS software [41] was used to fit the formula. Finally, we obtained parameter a = 23.989, parameter c = 0.004, and model determination coefficient R^2 = 0.870, standard error (SE) = 0.942.

In the established site index model, one group of data represents the dominant tree height and age of the sub-compartments, and the other group represents the standard age (the age at which height the growth of Chinese fir stands becomes stable) and the SI of Chinese fir. Relevant studies have shown that the standard age T of Chinese fir is 20 years [40]. Hence, we set $HT = HT_2$, $t = t_2$, SI = HT_1, and $T = t_1 = 20$, and placed each parameter into Equation (2). After transformation, the SI model of Chinese fir is shown as follows:

$$\mathrm{SI} = 23.989\left(\frac{HT}{23.989}\right)^{\frac{ln(1-exp(-0.004t))}{ln(1-exp(-0.004*20))}} \quad (3)$$

(2) Site grade division

According to Equation (3), the SI of each Chinese fir sub-compartment can be calculated, and the frequency of each sub-compartment of SI is statistically shown in Table 2:

Table 2. The SI frequency distribution of the Chinese fir sub-compartments.

SI	Sub-Compartment Frequency	SI Grade	SI Frequency
6	86	Grade III	907
8	187		
10	634		
12	530	Grade II	874
14	344		
16	106	Grade I	122
18	11		
20	5		

Table 2 shows eight classes of site indices in the Chinese fir sub-compartments in Lin'an District, with the indices ranging from 6 to 20. The distribution of sub-compartments under different site indices was unbalanced, among which the number of sub-compartments with site indices 10, 12, and 14 was high, accounting for 79.2% of the total, while the number of sub-compartments with other site indices was low. Given that numerous site indices would affect the accuracy of the final classification model, the site grade of the sub-compartment in this study was divided into three categories based on the frequency distribution of sub-compartments in the data: grades 16–20 are Grade I, and the frequency of sub-compartments is 122, which is regarded as an excellent site quality grade; grades 12–14 are Grade II, and the frequency of sub-compartments is 874, which is considered medium site quality grade; and grades 6–10 are Grade III, and the frequency of sub-compartments is 907, which is regarded as inferior site quality grade.

2.2.2. Data Preprocessing

(1) Discretization of continuous data

The discretization of continuous variables will improve the prediction accuracy of the random forest model [42]. In this research data, the altitude, undergrowth vegetation height, undergrowth vegetation coverage, and canopy closure belonged to continuous variables. In this study, the four factors were discretized by referring to the grading standards of some factors specified in the "Technical Operation Rules for Forest Resources Planning and Design Survey of Zhejiang Province (2014 Edition)" and the distribution of data in this study. The results are shown in Table 3.

Table 3. The discretization of continuous site factors.

Site Factors	Discrete Classification Standard
Altitude	High: ≥1000 m; medium: 500–1000 m; low: <500 m
Undergrowth vegetation height	High: ≥60 cm; medium: 30–60 cm; low: <30 m
Undergrowth vegetation coverage	High: ≥60%; medium: 30%–60%; low: <30%
Canopy closure	High: ≥70%; medium: 40%–70%; low: <40%

(2) Balanced sampling plans

Table 2 shows that approximately 93.6% of the Chinese fir sub-compartments in Lin'an District were in the middle or low site grade, while the number of the excellent site classes only accounted for about 6.4% of all sub-compartments. The uneven frequency distribution of the site-level data would affect the performance of the model. Therefore, the sample data in this study were over-sampled by the SMOTE algorithm. The SMOTE algorithm is an oversampling method widely used in the classification of data imbalance [43]. The principle of this algorithm is to synthesize a new minority class sample. That is, for each minority class sample Xi, a sample Xij was randomly selected from the k-nearest neighbor. Moreover, a point on the line between Xi and Xij was randomly selected as a newly synthesized minority class sample.

The SMOTE algorithm principle is explained in Figure 2.

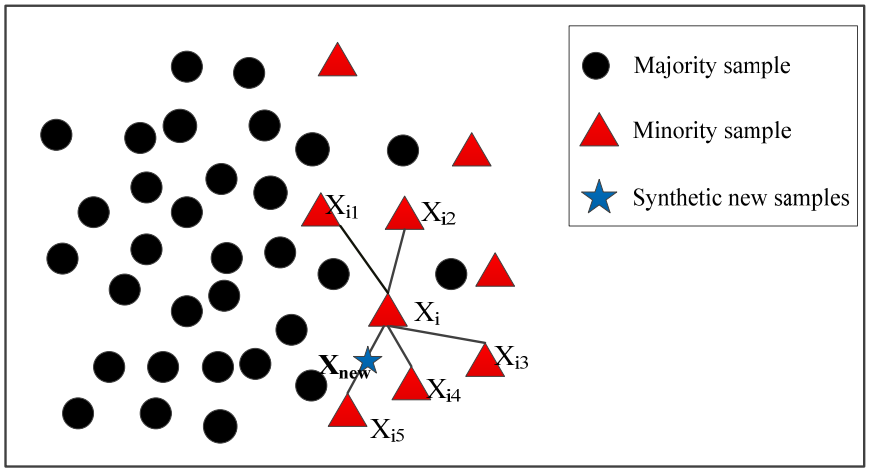


Figure 2. The SMOTE algorithm principle based on k proximity.

The algorithm can be implemented through the third-party library "imblear.over_sampling" of Python, and the number of minority class samples would be a multiple of the original sample number after balancing. In this study, the number of each SI grade after the SMOTE algorithm balance was 907 (see Table 4) so as to not lose the original number of majority samples.

Table 4. The number of samples before and after data balance.

Sample Types	SI Grades		
	Grade I	**Grade II**	**Grade III**
Original sample	122	874	907
Balanced sample	907	907	907

Given that the newly added samples were generated according to distance through the original samples, non-integer data cannot be avoided in the newly added samples, thereby affecting the construction of the subsequent classification model. Therefore, the newly generated sample numbers were rounded to integers after oversampling.

2.2.3. Site Factor Reduction Based on Rough Set

In this study, Python was used to construct a factor reduction algorithm proposed by Pawlak [18]. Site factors and SI grades were taken as input variables and decision variables, respectively. First, the core attributes of the site factors relative to site grades was calculated using the elimination method. The initial reduction table was composed of core attributes, and the importance degree of other remaining attributes other than nuclear attributes was calculated each time. Moreover, the attribute with the largest value of importance was selected and added to the reduction set until the importance degree of all of the remaining attributes was 0. That is, the value of the system's dependency function does not change when any new attribute is added.

The main calculation process is as follows.

Input: Attribute reduction decision table T = (U,C,D), where U represents the dataset of Chinese fir sub-compartments, C is the conditional attribute in dataset U (i.e., each site factor), and D is the SI grade.

Output: Attribute reduction set R.

Step 1: Calculate the positive domain POSc(D) of the site grade for site factors, select a random attribute i from set C of the site factors, and calculate POSc-{ci}(D) without attribute i.

Step 2: If POSc(D) is equal to POSc-{ci}(D), then the remaining site factor Cj is randomly selected on the basis of C-Ci. Compare whether or not POSc(D) is equal to POSc-{ci,cj(i ≠ j)}(D). If they are still equal, then continue to remove the remaining site factors until the attribute set that is not equal to POSc(D) is found. COREc(D), so the core attribute of the site factors dataset C relative to SI grade D can be obtained.

Step 3: Take the core attributes in Step 2 as the initial reduction attribute set R.

Step 4: On the basis of core attributes R, the importance of each non-core attribute C_k to decision set D was calculated, and the calculation formula is as follows:

$$\mathrm{sig}(C_k RD) = \frac{\left|POS_{R\cup\{C_k\}}\right| - |POS_R(D)|}{|U|} \tag{4}$$

where sig(C_k,R,D) is the importance of attribute C_k to SI grade and other variables are as previously defined.

Step 5: Take the most important attribute C_l and add this attribute to R, which is $R = R \cup C_l$

Step 6: Return to Step 4 and loop until the condition is not satisfied.

Rationality of the reduced attributes of the rough set can be expressed according to dependence degree e, which is expressed as follows:

$$e = \frac{|POS_C(D)|}{|U|} \tag{5}$$

where *POSc(D)* is the positive domain of site quality grades for site factors; U represents the entire dataset; and e is typically between 0 and 1. Moreover, e reflects the ability to correctly classify sub-compartments into corresponding site quality levels according to reduced attributes. The larger the value of e, the more the reduced attributes can explain the final classification results. The related concepts and specific derivation process of rough set theory can be referred to in [18].

2.2.4. Site Classification Modeling of Random Forest

(1) Random forest principle

Random forest is an integrated learning algorithm based on decision tree proposed by Breiman [44]. This algorithm uses bootstrap re-sampling to extract multiple samples from the original data, conducts decision tree modeling for each bootstrap sample, and obtains the classification results of the optimal decision tree through voting. The modeling steps are as follows.

Step 1: k samples are extracted from the original training set D to construct the D1, D2 . . . Dk sub-training set. The amount of data in the sub-training set is the same as that of the original training set.

Step 2: k decision trees are constructed using k sub-training sets. In splitting the decision tree, a certain feature was randomly selected from all features (site factors), and the optimal feature was selected thereafter for segmentation.

Step 3: Test sets were used for prediction, and the k classification results were obtained from k decision trees.

Step 4: The final classification result was obtained by voting on the k classification results.

The construction process of the random forest model is shown in Figure 3.

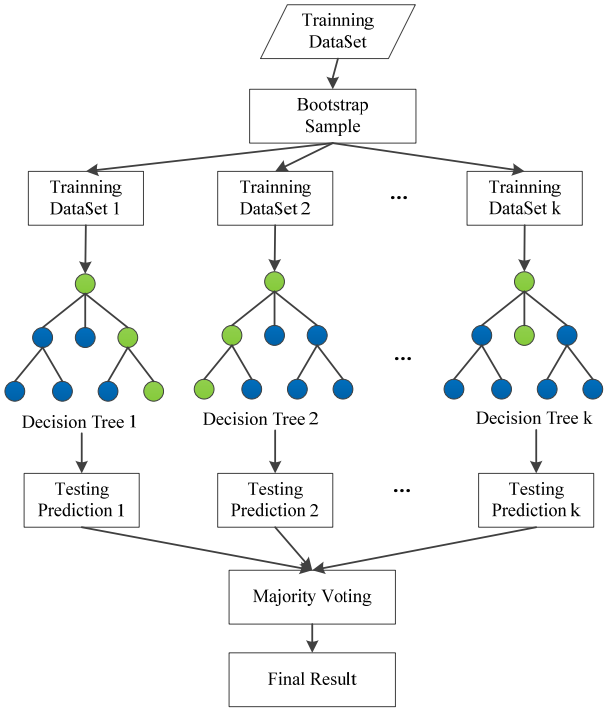
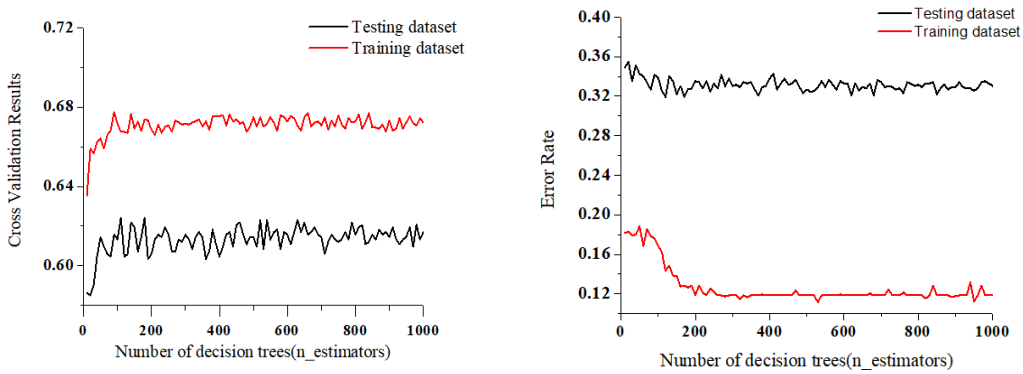


Figure 3. The principles of random forest modeling.

(2) Implementation of the random forest model

This study used Python to build a random forest model on the Visual Studio Code platform [45]. Python provides a program package "RandomForestClassifier" to build the random forest model. The random forest model has two important customized parameters: "n_estimators" and "max_features". In particular, "n_estimators" is the number of sub-models. In general, the more the number of sub-models, the better the performance of the model and the more stable the accuracy of prediction. However, this situation may slow down the calculation process. Some studies have shown that the final value of "n_estimators" only needs to meet the requirement that the overall error of the random forest tends to be stable [46]. This study set the "n_estimators" range between 10 and 1000, and 10 was taken as the step to calculate the accuracy of the model cross validation and overall error of the model. The results are shown in Figure 4a,b.



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

Figure 4. The relationship between model parameter "n_estimators" and model performance. (**a**) Cross-validation results and number of sub-models. (**b**) Model error rate and number of sub-models.

Figure 4a,b shows that when the number of sub-models was over 200, the cross validation results and error rate tend to be stable. Therefore, the "n_estimators" value in this model was set to 200.

"Max_features" refers to the maximum number of features involved in the judgment of node splitting. In general, the smaller the "max_features", the more different the trees in the random forest will be. However, if the "max_features" are considerably small (set to 1), choosing the feature that can be tested in division would be impossible. In the Python-based random forest classification model library "sklearn," the recommended value for "max_features" is the square root of the total number of features [47]. Therefore, the "max_features" in this model was the square root of the total number of site factors, which was the square root of 20 and the square root of the number of site factors after reduction.

(3) Model evaluation method

In classification problems, a confusion matrix is generally used to display the prediction of each category, typically based on the following four indicators.

True Positive (TP): The number of sub-compartments of the real category of the sub-compartment is in an SI grade, and its prediction category is also in that SI grade.

False Positive (FP): The number of sub-compartments of the real category of the sub-compartment is not in an SI grade, but its prediction category is in that SI grade.

True Negative (TN): The number of sub-compartments of the real category of the sub-compartment is not in an SI grade, and its prediction category is also not in that SI grade.

False Negative (FN): The number of sub-compartments of the real category of the sub-compartment is in an SI grade, but its prediction category is not in that SI grade.

According to the confusion matrix, various indicators can be derived to measure the performance of the model including precision, recall, and accuracy. These indicators are calculated as follows:

Precision = TP/TP + FP;

Recall = TP/TP + FN; and

Accuracy = (TP + TN)/(TP + TN + FP + FN).

In this study, the experimental data were divided into the training and test samples in a 7:3 ratio. That is, training samples were used to build the model and test samples were used to test the accuracy of the model.

(4) Importance evaluation of variables

Random forest has the function of feature importance assessment, and the importance of each factor is often measured according to the reduction in the classification accuracy. In random forest, about a 1/3 of data are not selected in each sampling (i.e., out-of-bag (OOB) data). The OOB error Er of the rth subtree in the model was calculated according to the OOB data. Thereafter, the OOB error of the rth tree was calculated again by adding feature j, denoted as Er_j. The importance M_j of feature j can be expressed as follows:

$$M_j = \sum_{r=1}^{N} Er_j - Er \tag{6}$$

where N is the number of random forest subtrees and other variables are as previously defined. M_j is normalized to 0–1; the greater the value, the greater the importance of this feature.

3. Results

3.1. Attribute Reduction Results Based on Rough Sets

Table 5 shows that seven of the 20 site factor attributes were reduced: forest protection grade, soil texture, altitude, land type, soil types, landform, and age group. These factors can be disregarded because they cannot determine the site grade of Chinese fir sub-compartments. The following 13 attributes were retained: naturalness, stand origin, plant

community structure, forest class, soil layer thickness, humus layer thickness, undergrowth vegetation coverage, undergrowth vegetation height, undergrowth vegetation species, slope position, slope gradient, slope direction, and canopy closure. Among the retained attributes, canopy closure, slope direction, slope gradient, slope position, undergrowth vegetation species, undergrowth vegetation height, and undergrowth vegetation coverage were the essential attributes of site classification (i.e., core attributes).

Table 5. The attribute reduction results of the rough set.

Categories	Specific Site Factors	Factor Numbers	Dependence Degree *e*
Reduced attributes	Forest protection grade, soil texture, altitude, land type, soil types, landform, age group Naturalness, stand origin, plant community structure, forest class, soil layer thickness, humus layer thickness, undergrowth vegetation coverage, undergrowth vegetation height, undergrowth vegetation species, slope position, slope gradient, slope direction, canopy closure	7	0.94
Reserved attributes		13	
Core attributes	Canopy closure, slope direction, slope gradient, slope position, undergrowth vegetation species, undergrowth vegetation height, undergrowth vegetation coverage	7	

Classification dependence degree e of the rough set was 0.94, which was close to 1. The results showed that reserved attributes can reasonably explain the classification results of Chinese fir sub-compartments.

3.2. Results of Classification Model Based on Random Forest

3.2.1. Comparison of Model Accuracy

Two 3 × 3 multi-classification confusion matrices were obtained, as shown in Figure 5. In the training dataset, scheme A without an attribute reduction and scheme B with an attribute reduction were compared. Moreover, there were 639 sub-compartments with SI Grade I, 508 of which were correctly classified as SI Grade I forestland in Scheme A, and the recall rate was 79.50%. A total of 564 were in Scheme B, and the recall rate was 88.26%. Meanwhile, there were 628 sub-compartments with Grade II site quality, 510 of which were correctly classified as Grade II forestland in Scheme A, with a recall rate of 81.21%. A total of 550 were correctly classified as Grade II forestland in Scheme B, with a recall rate of 87.58%. Moreover, there were 637 sub-compartments with Grade III site quality, 510 of which were correctly classified as Class III forestland for Scheme A, with the calculated recall rate of 80.06%. A total of 575 were correctly classified as Class III forestland for Scheme B, with a recall rate of 90.27%.

Predicted

Actual	I	II	III	Total
I	508	84	47	639
II	66	510	52	628
III	53	74	510	637
Total	627	668	609	1904

(**a**)

Predicted

Actual	I	II	III	Total
I	564	40	35	639
II	46	550	32	628
III	31	31	575	637
Total	641	621	642	1904

(**b**)

Figure 5. The confusion matrices with training data. (**a**) Scheme A confusion matrix with unreduced attributes. (**b**) Scheme B confusion matrix with reduced attributes.

The model was validated with the test data, and the confusion matrices based on the test data are shown in Figure 6. Confusion matrices of the two schemes were compared on the test data. In particular, there were 268 sub-compartments with Grade I site quality, 147 of which were correctly classified as Grade I forestland for Scheme A, with a calculated recall rate of 54.85%. A total of 169 were for Scheme B, with a recall rate of 63.06%. Meanwhile, there were 279 sub-compartments with the site quality of Grade II, 131 of which were correctly classified as Grade II forestland in scheme A, with a recall rate of 46.95%. A total of 140 were correctly classified as Grade II forestland in Scheme B, with a recall rate of 50.18%. Finally, there were 270 sub-compartments with the site quality of Grade III, 195 of which were correctly classified as Class III forestland in Scheme A, with a calculated recall rate of 72.22%. A total of 244 were correctly classified as class III forestland in Scheme B, with a recall rate of 90.37%.

Predicted

Actual	I	II	III	Total
I	147	99	22	268
II	112	131	36	279
III	31	44	195	270
Total	290	274	253	817

(**a**)

Predicted

Actual	I	II	III	Total
I	169	85	14	268
II	116	140	23	279
III	11	15	244	270
Total	296	240	281	817

(**b**)

Figure 6. The confusion matrices with testing data. (**a**) Scheme A confusion matrix with unreduced attributes. (**b**) Scheme B confusion matrix with reduced attributes.

Table 6 shows the classification level of models with the two schemes. Compared with scheme A, Scheme B reduced seven site factor attributes, accounting for 35.0% of the total, and the modeling time was reduced by 2.71 s, accounting for 50.19% of the total time of Scheme A. In the training data, the accuracy of the model in Scheme A was 80.37%, the recall was 80.26%, and the accuracy was 86.83%. The accuracy of the model in Scheme B was 88.70%, the recall was 88.70%, and the accuracy was 92.47%. In the test data, the model accuracy in Scheme A was 58.52%, the recall was 58.00%, and the accuracy was 71.93%. The model accuracy in Scheme B was 67.42%, the recall was 67.87%, and the accuracy was 78.46%.

Table 6. A comparison of the two schemes for the site quality grade evaluation of Chinese fir.

Schemes	Number of Factors	Training Time	Training Dataset			Testing Dataset		
			Precision	Recall	Accuracy	Precision	Recall	Accuracy
Scheme A	20	5.40 s	0.8037	0.8026	0.8683	0.5852	0.5800	0.7193
Scheme B	13	2.69 s	0.8870	0.8870	0.9247	0.6742	0.6787	0.7846

Note that rough sets played an obvious role in the simplified classification, as shown in Table 6. After attribute reduction, the training time and complexity of the model were shorter and reduced, respectively. Moreover, the accuracy, recall, and accuracy of the model after attribute reduction were also improved, making the model considerably valuable for promotion.

3.2.2. Application of the Model

The model combining the rough set and random forest was applied to the quality evaluation of Chinese fir in the Lin'an District. Data of 312 Chinese fir sub-compartments

that were not used in the modeling and 1232 non-forestland sub-compartments were selected. A total of 13 site factor attributes of the sub-compartments were input into the established site quality classification model. Finally, the site grade of the sub-compartments could be output and visualized by GIS. The site grade of a Chinese fir sub-compartment is clearly presented in Figure 7, which is convenient for forestry workers to divide Chinese fir forestland and choose thinning measures according to the principle of "matching tree species with site".

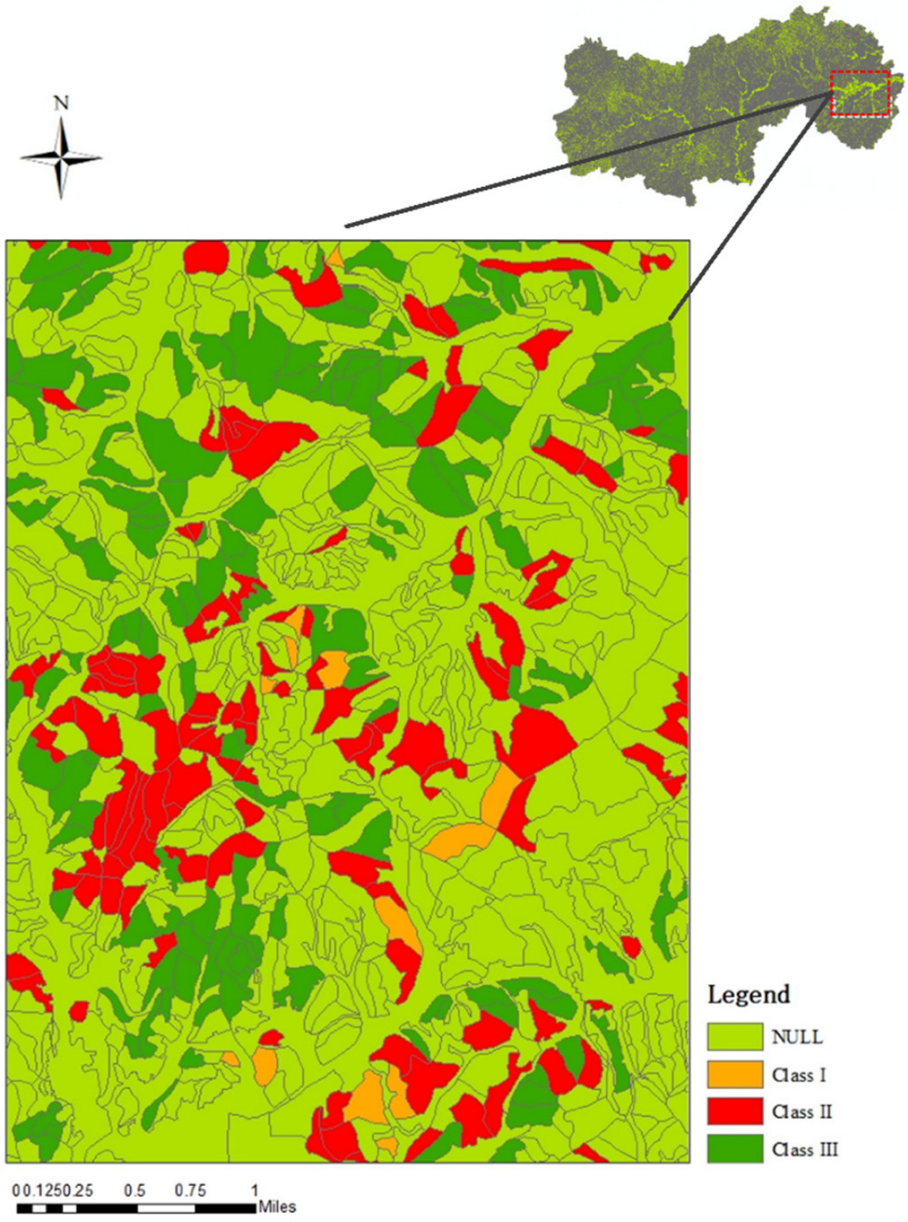


Figure 7. The site quality grade prediction of a Chinese fir plantation in Lin'an District.

3.3. Importance Assessment of Site Factors

The importance of 13 site factors was evaluated using the evaluation method of feature importance (in Section 2.2.4) to analyze their influence on the growth of Chinese fir. Importance values were normalized from 0 to 1 and arranged from high to low, as shown in Figure 8.

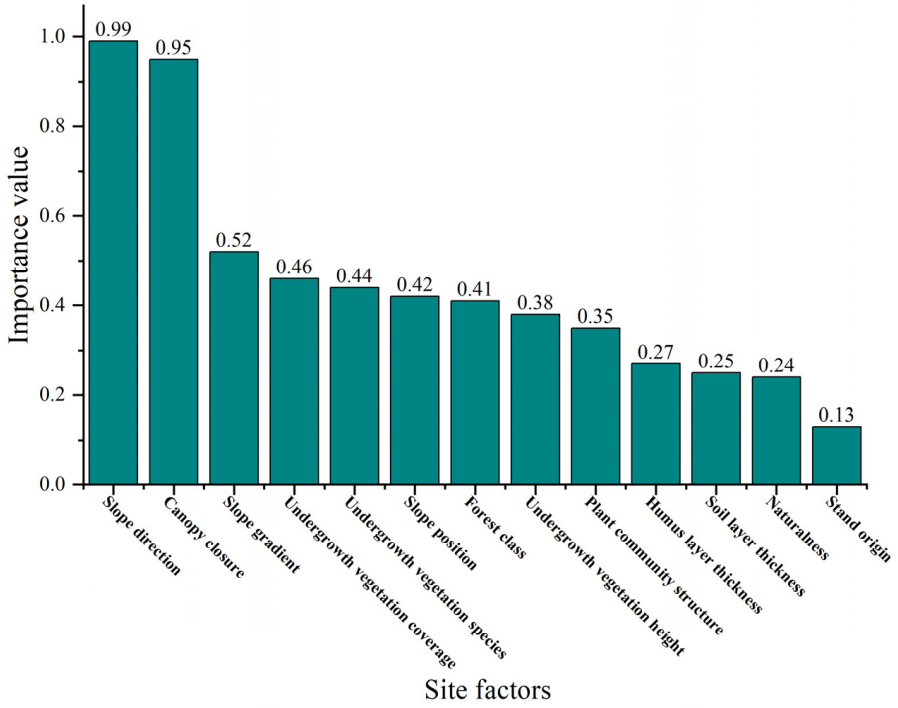


Figure 8. The importance ranking of the site factors.

Figure 8 shows that among the 13 site factors, the slope direction and canopy closure had the greatest influence on the growth of Chinese fir with influence values of 0.99 and 0.95, respectively, followed by the slope gradient with an influence value of 0.52. Factors with minimal influence on the growth of Chinese fir were the humus layer thickness, soil layer thickness, naturalness, and stand origin, with importance values below 0.3.

4. Discussion

The results of the rough set study showed that the main factors affecting the growth of Chinese fir were naturalness, stand origin, plant community structure, forest class, soil layer thickness, humus layer thickness, undergrowth vegetation coverage, undergrowth vegetation height, undergrowth vegetation species, slope position, slope gradient, slope direction, and canopy closure. These factors played a key role in the site quality classification of the Chinese fir sub-compartments. In the attribute reduction based on rough set, Pawlak indicated that the reduction algorithm is suitable for dealing with discrete variables. However, some forestry data belonged to continuous data, in which the Pawlak algorithm was introduced to process this type of data [18]; continuous data were often converted into discrete data, inevitably resulting in information loss [48]. To solve this problem, fuzzy rough set, similar relation rough set, and neighborhood relation models can be introduced to study the attribute reduction of site factors in subsequent research.

The results of the site quality classification model based on random forest showed that this model, based on reduced attributes, was more simplified and the model training efficiency was higher. The accuracy, recall rate, and accuracy of the model were relatively improved in the training and testing sets compared with the model without attribute reduction. The random forest model is an extension of the decision tree model. Chen et al. once used the decision tree to construct the quality classification model of Chinese fir, and her research results showed that the classification accuracy of the model was lower than that of the random forest model [15]. At present, some scholars have used random forest algorithm to evaluate site quality, but in the selection of site factors, almost all of them were subjective selection based on experience [49], and the rough set in this study could well solve the subjective problem of site factor selection. Moreover, the effects of 13 site factors on the growth of Chinese fir were analyzed using the variable importance assessment

function of the random forest model. The results showed that the most influential site factors were slope direction, canopy closure, and slope gradient, while the less influential factors were the humus layer thickness, soil layer thickness, naturalness, and stand origin in the study area. The reasons were as follows. The change in the slope direction and slope gradient have a certain influence on solar radiation, soil fertility, and air temperature. Hence, the slope direction and slope gradient have immense influence on the growth of Chinese fir.

Related studies in the same region have shown that the steeper the slope, the worse the stand quality. The reason is that the slope has an impact on the microclimate of the stand. The place where the slope is considerably steep is often located in the windward with the thinner soil layer, which is not conducive to the growth of Chinese fir [50]. Some studies have also shown that Chinese fir on the northeast and northwest slopes has better site quality than that on the south slope, indicating that Chinese fir is more suitable for growing on shady or semi-shady slopes [5]. Some studies have also shown that site factors have different effects on the growth of Chinese fir in different growth stages of stands. Slope position is the main factor affecting the growth of Chinese fir in young and middle-age stands, while humus thickness is the most critical factor affecting the growth of Chinese fir in near-mature and over-mature stands [16].Canopy closure is the embodiment of stand density. The change in canopy closure can indirectly affect changes in solar radiation, stand air humidity, growth environment of undergrowth vegetation, soil physical and chemical conditions, and the types and activity intensity of microorganisms in the soil. Some studies have also shown that the density of the stand indirectly affects the site conditions of vegetation [51]. Therefore, canopy closure is an important factor affecting the growth of Chinese fir. Although there are many studies that have indicated that soil layer thickness and humus layer thickness have relatively important effects on soil quality [52–54], there are a few types of soil in the planting area of Chinese fir in Lin'an District, most of which are yellow and red soils. In addition, most of the soil is thick, so the influence of soil thickness and humus layer thickness on the growth of Chinese fir is not evident. At present, no study has been conducted to analyze the impact of naturalness and origin on site quality. The data in this study indicated that stands with naturalness Class III accounted for 94.5%, and the rest of the stands with naturalness II and I merely accounted for 5.5%. In the origin of stand, plantations accounted for 96.0% and natural forest only accounted for 4.0%. Thus, the imbalance in the experimental data was also the main reason that the preceding factors did not clearly classify the site quality of Chinese fir. Furthermore, plantations are probably located in specific (pre-selected) locations, therefore, the site factors affecting the growth of Chinese fir showed inconsistent conclusions with other references [55,56], the results of this study are only applicable to the study area, so we still need to verify the applicability of the results to other regions.

The results of this study proved that the method of forest site quality evaluation combined with rough set and random forest could deal well with the nonlinear relationship between forest site quality and site factors as well as overcome the limitation and subjectivity of the artificial selection of site factors. The random forest model can improve the accuracy of classification and prediction without significantly increasing the amount of computation. In the model construction, there are few adjustment parameters, and it can also be used to evaluate the importance of features. In general, the model has numerous advantages in classification. The model can predict the site quality of Chinese fir with the site factors and also judge whether or not there is forestland suitable for the growth of Chinese fir. Meanwhile, the algorithm was edited using Python, which has strong universality and compatibility. Finally, the programs proposed in this study can be used on different software platforms, thereby providing a new idea for the application of big data in Chinese forestry.

The main innovation of this study was to apply the rough set theory and random forest model to the problem of "matching tree species with site" with satisfactory results. In future research, we should deeply analyze the impact of each site factor on the site quality

of the stand and the interaction between site factors under different climates, environments, stand ages, and stand densities, so the growth environment of Chinese fir is in the best combination state to achieve the best productivity. In addition, the random forest model has potential wide application. The proposed model was only for Chinese fir species, and random forest models of other species can be established in the future. Future models can provide scientific theoretical basis for further discussion of the spatial distribution of the forest site quality grade and forest land utilization planning, and provide technical support for improving forestry information management.

5. Conclusions

This paper studied the site quality classification model of Chinese fir in Lin'an District, Zhejiang Province. First, an SI model was constructed using ADA to divide the site grade of the Chinese fir sub-compartments in the study area. The original data were discretized and balanced to improve the accuracy of the subsequent models. Thereafter, 20 site factors were selected as the initial factors, which were reduced using rough set theory. Eventually, 13 site factors closely related to the site quality of Chinese fir in the study area were obtained. The random forest model of machine learning was introduced into this study, and site quality classification models based on the initial and reduction factors were constructed. This study proposed a complete process of data processing, modeling, and evaluation. Moreover, the optimal model was used to classify the site quality grade of Chinese fir sub-compartments that were not used in the modeling, and the classification results were visualized. Finally, the importance of factors affecting the site quality of Chinese fir was evaluated.

Author Contributions: Conceptualization, C.D.; Methodology, C.D. and Y.C.; Validation, C.D.; Formal analysis, C.D. and Y.C.; Resources, X.L.; Data curation, Z.M.; Writing—original draft preparation, C.D.; Writing—review and editing, C.D.; Visualization, C.D. and J.B.; Supervision, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Zhejiang Province (grant number LQ21C160004) and the Research Development Fund Project of Zhejiang Agriculture & Forestry University (grant number 203402010801).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available on request from the corresponding author.

Acknowledgments: We are grateful to Kai Xia at Zhejiang A&F University for supplying the valuable model data. We would like to express our gratitude to the English language editors, who helped in checking the grammar mistakes.

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

References

1. Yang, H. A Study of Site Quality Evaluation of Chinese Fir Plantation Based on NFI Data of Zhejiang Province. Master's Thesis, Zhejiang A & F University, Hangzhou, China, 2019.
2. Carmean, W.H. Forest Site-Quality Estimation Using Forest Ecosystem Classification in Northwestern Ontario. *Environ. Monit. Assess.* **1996**, *39*, 493–508. [CrossRef]
3. Skovsgaard, J.P.; Vanclay, J.K. Forest site productivity: A review of evolution of dendrometric concepts for even-aged stands. *Forestry* **2008**, *81*, 13–31. [CrossRef]
4. Quichimbo, P.; Jiménez, L.; Veintimilla, D.; Tischer, A.; Günter, S.; Mosandl, R.; Hamer, U. Forest Site Classification in the Southern Andean Region of Ecuador: A Case Study of Pine Plantations to Collect a Base of Soil Attributes. *Forests* **2017**, *8*, 473. [CrossRef]
5. Dong, C.; Fang, L. Association analysis between the site index Model and the Site Factors of *Cunninghamia Lanceolata* Timber Forest in Western Zhejiang Province. *Nat. Environ. Pollut. Technol.* **2019**, *18*, 359–368.
6. Qiu, H.; Liu, S.; Zhang, Y.; Li, J. Variation in height-diameter allometry of ponderosa pine along competition, climate, and species diversity gradients in the western United States. *For. Ecol. Manag.* **2021**, *497*, 119477. [CrossRef]
7. Andrés, B.O.; Clemente, G.A.; Miren-del, R.; Montero, G. Regional changes of Pinus pinaster site index in Spain using a climate-based dominant height model. *Can. J. For. Res.* **2010**, *40*, 2036–2048.

8. Hlásny, T.; Trombik, J.; Bošeľa, M.; Merganič, J.; Marušák, R.; Šebeň, V.; Štěpánek, P.; Kubišta, J.; Trnka, M. Climatic drivers of forest productivity in Central Europe. *Agric. For. Meteorol.* **2017**, *234–235*, 258–273. [CrossRef]
9. Gao, R.; Xie, Y.; Lei, X.; Lu, Y.; Su, X. Study on prediction of natural forest productivity based on random forest model. *J. Cent. South Univ. For. Technol.* **2019**, *39*, 39–46.
10. Site Productivity and Forest Growth Modelling Strategies: Monospecific Versus Mixed Species Forests. Available online: https://www.researchgate.net/publication/344201149 (accessed on 26 December 2019).
11. Novor, S.; Abugre, S. Growth Performance, Undergrowth Diversity and Carbon Sequestration Potentials of Tree Species Stand Combinations, Ghana. *Open J. For.* **2020**, *10*, 135–154. [CrossRef]
12. Eslamdoust, J.; Sohrabi, H. Carbon storage in biomass, litter, and soil of different native and introduced fast-growing tree plantations in the South Caspian Sea. *J. For. Res.* **2018**, *29*, 449–457. [CrossRef]
13. Huang, S.; Ramirez, C.; Conway, S.; Kennedy, K.; Kohler, T.; Liu, J. Mapping site index and volume increment from forest inventory, Landsat, and ecological variables in Tahoe National Forest, California, USA. *Can. J. For. Res.* **2017**, *47*, 147–156. [CrossRef]
14. Kang, S.-p.; Kim, J.-y.; Ahn, K.-w. Site Index Estimation and Suitable-Land Evaluation of *Cryptomeria japonica* and *Chamaecyparis Obtusa*—Focused on Jeju Special Self-Governing Province and Southern Regions. *TJOKI* **2015**, *27*, 125–144.
15. Chen, Y.; Wu, B.; Qi, Y. Using Machine Learning to Assess Site Suitability for Afforestation with Particular Species. *Forests* **2019**, *10*, 739. [CrossRef]
16. Guo, Y.; Liu, Y.; Wu, B. Evaluating Dividing Rank and Quantification of Site Quality of Suitable Land for Forest in Fujian Province, China. *J. Northeast For. Univ.* **2014**, *42*, 54–59.
17. Lv, F.Z.; Luo, H.J.; Lv, Y. Study on forestland quality indexes and their application. *J. For. Environ.* **2015**, *1*, 87–91.
18. Pawlak, Z. Rough set theory and its applications to data analysis. *Cybern. Syst.* **1998**, *29*, 611–688. [CrossRef]
19. Vanclay, J.K.; Henry, N.B. Assessing Site Productivity of Indigenous *Cypress Pine* Forest in Southern Queensland. *Emp. For. Rev.* **1988**, *67*, 53–64.
20. Vanclay, J.K. Assessing site productivity in tropical moist forests: A review. *For. Ecol. Manag.* **1992**, *54*, 257–287. [CrossRef]
21. Zhu, G.Y.; Hu, S.; Chhin, S.; Zhang, X.Q.; He, P. Modelling site site index of Chinese fir plantations using a random effects model across regional site types in Hunan province, China. *For. Ecol. Manag.* **2019**, *446*, 143–150. [CrossRef]
22. Duan, G.; Wang, Q.; Fu, L. Comparison of Different Height–Diameter Modelling Techniques for Prediction of Site Productivity in Natural Uneven-Aged Pure Stands. *Forests* **2018**, *9*, 63. [CrossRef]
23. Stankova, T.; Diéguez-Aranda, U. A tentative dynamic site index model for Scots pine (*Pinus sylvestris*) plantations in Bulgaria. *Silva Balc.* **2012**, *13*, 5–17.
24. Trim, K.R.; Coble, D.W.; Weng, Y.H.; Stovall, J.P.; Hung, I.K. A New Site Index Model for Intensively Managed Loblolly Pine (*Pinus taeda*) Plantations in the West Gulf Coastal Plain. *For. Sci.* **2019**, *66*, 2–13. [CrossRef]
25. Socha, J.; Tyminska-Czabanska, L.; Grabska, E.; Orzel, S. Site Index Models for Main Forest-Forming Tree Species in Poland. *Forests* **2020**, *11*, 301. [CrossRef]
26. Batho, A.; García, O. A Site Index Model for Lodgepole Pine in British Columbia. *For. Sci.* **2014**, *60*, 982–987. [CrossRef]
27. Daniel, M.; Juan, G.; Roque, R. National-scale assessment of forest site productivity in Spain. *For. Ecol. Manag.* **2018**, *417*, 197–207.
28. Kahriman, A.; Yavuz, H.; Ercanli, I. Site index conversion equations for mixed stands of Scots pine(*Pinus sylvestris* L.) and Oriental beech (*Fagus orientalis Lipsky*) in the Black Sea Region, Turkey. *Turk. J. Agric. For.* **2013**, *37*, 488–494. [CrossRef]
29. Bayat, M.; Bettinger, P.; Hassani, M.; Heidari, S. Ten-year estimation of Oriental beech (*Fagus orientalis Lipsky*) volume increment in natural forests: A comparison of an artificial neural networks model, multiple linear regression and actual increment. *Forestry* **2021**, *94*, 598–609. [CrossRef]
30. Shen, J. Study on Site Quality Evaluation Methods of Uneven-Aged Coniferous and Broad-Leaved Mixed Stands in Guangdong Province. Ph.D. Thesis, Chinese Academy of Forestry, Beijing, China, 2018.
31. Liu, Z.L.; Peng, C.H.; Work, T.; Candau, J.N.; DesRochers, A.; Kneeshaw, D. Application of machine-learning methods in forest ecology: Recent progress and future challenges. *Environ. Rev.* **2018**, *26*, 339–350. [CrossRef]
32. Weng, Y.; Grogan, J.; Cheema, B.; Tao, J.; Lou, X.; Burkhart, H. Model-Based Growth Comparisons between Loblolly and Slash Pine and between Silvicultural Intensities in East Texas. *Forests* **2021**, *12*, 1611. [CrossRef]
33. Lou, X.; Weng, Y.; Fang, L.; Gao, H.; Grogan, J.; Hung, I.K.; Oswald, B.P. Predicting stand attributes of loblolly pine in West Gulf Coastal Plain using gradient boosting and random forests. *Can. J. For. Res.* **2020**, *51*, 807–816. [CrossRef]
34. Wang, Z.; Zhang, X.; Chhin, S.; Zhang, J.; Duan, A. Disentangling the effects of stand and climatic variables on forest productivity of Chinese fir plantations in subtropical China using a random forest algorithm. *Agric. For. Meteorol.* **2021**, *304*, 108412. [CrossRef]
35. Zhao, M.; Xiang, W.; Peng, C.; Tian, D. Simulating age-related changes in carbon storage and allocation in a Chinese fir plantation growing in southern China using the 3-PG model. *For. Ecol. Manag.* **2009**, *257*, 1520–1531. [CrossRef]
36. Li, Y.; Zhang, B.; Qin, S. Review of research and application of forest canopy closure and its measuring methods. *World For. Res.* **2008**, *21*, 40–46.
37. Guo, Y.; Wu, B.; Liu, Y. Research progress of site quality evaluation. *World For. Res.* **2012**, *25*, 47–52.
38. Chen, Y. Research on Matching Tree Species with Site and Growth Yield Benefit Assessment of Plantation-in the Case of *Cunninghamia lanceolata* and *Pinus massoniana* in Guizhou Province. Ph.D. Thesis, Beijing Forestry University, Beijing, China, 2020.

39. Zobel, J.M.; Schubert, M.R.; Granger, J.J. Shortleaf Pine (*Pinus echinata*) Site Index Equation for the Cumberland Plateau, USA. *For. Sci.* **2022**, *68*, 259–269. [CrossRef]
40. Guo, Y.; Han, Y.; Wu, B. Study on modelling of site quality evaluation and its dynamic update technology for plantation forests. *Nat. Environ. Pollut. Technol.* **2013**, *12*, 591–597.
41. Feng, L. *Principle of Regression Analysis Method and Practical Operation of SPSS*; China Finance Publishing House: Beijing, China, 2004; pp. 32–46.
42. Ye, Y.; Wu, Q.; Huang, J.Z.; Ng, M.K.; Li, X. Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognit.* **2013**, *46*, 769–787. [CrossRef]
43. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
44. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
45. Visual Studio Code Builds the Python Development Environment. Available online: https://www.cnblogs.com/liangqihui/articles/9241597.html (accessed on 29 June 2018).
46. Ziegler, A.; Konig, I.R. Mining data with random forests: Current options for real-world applications. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2014**, *4*, 55–63. [CrossRef]
47. Yoo, J.E. Random forests, an alternative data mining technique to decision tree. *J. Educ. Eval.* **2015**, *28*, 427–448.
48. Jensen, R.; Shen, Q. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 1457–1471. [CrossRef]
49. Wang, Y.; Feng, Z.; Ma, W. Analysis of Tree Species Suitability for Plantation Forests in Beijing (China) Using an Optimal Random Forest Algorithm. *Forests* **2022**, *13*, 820. [CrossRef]
50. Song, J. Discussion of the relationship between Chinese fir growth and environmental conditions. *West. China Technol.* **2008**, *7*, 54–55.
51. Watt, M.S.; Kimberley, M.O.; Dash, J.P.; Harrison, D. Spatial prediction of optimal final stand density for even-aged plantation forests using productivity indices. *Can. J. For. Res.* **2017**, *47*, 527–535. [CrossRef]
52. Sewerniak, P. Site index of Scots pine stands in south-western Poland in relation to forest site types and soil units. *SYLWAN* **2013**, *157*, 516–525.
53. Sacewicz, W.A.; Bijak, S. Effect of selected soil properties on site index of oak stands in the Miezyrzec Forest District. *SYLWAN* **2018**, *162*, 3–11.
54. Guner, S.T. Relationships between Site Index and Ecological Variables of Oriental Beech Forest in the Marmara Region of Turkey. *Fresenius Environ. Bull.* **2021**, *30*, 6920–6927.
55. Fang, J.; Shen, Z.; Cui, H. Ecological characteristics of mountains and research issues of mountain ecology. *Biodivers. Sci.* **2004**, *12*, 10–19.
56. Krner, C. The use of altitude in ecological research. *Trends Ecol. Evol.* **2007**, *22*, 569–574. [CrossRef]