remote sensing

MDPI

Article

Study on the Regeneration Probability of Understory Coniferous Saplings in the Liangshui Nature Reserve Based on Four Modeling Techniques

Haiping Zhao [1,2], Yuman Sun [1,2], Weiwei Jia [1,2,*], Fan Wang [1,2], Zipeng Zhao [1,2] and Simin Wu [1,2]

[1] Department of Forest Management, School of Forestry, Northeast Forestry University, Harbin 150040, China; zhaohp2021@nefu.edu.cn (H.Z.); symsfs@nefu.edu.cn (Y.S.); wangfan@nefu.edu.cn (F.W.); zip110194@nefu.edu.cn (Z.Z.); wusimin1017@nefu.edu.cn (S.W.)

[2] Key Laboratory of Sustainable Forest Ecosystem Management-Ministry of Education, School of Forestry, Northeast Forestry University, Harbin 150040, China

* Correspondence: jiaww@nefu.edu.cn; Tel.: +86-451-8219-1215

Abstract: Forests are one of the most important natural resources for humans, and understanding the regeneration probability of undergrowth in forests is very important for future forest spatial structure and forest management. In addition, the regeneration of understory saplings is a key process in the restoration of forest ecosystems. By studying the probability of sapling regeneration in forests, we can understand the impact of different stand factors and environmental factors on sapling regeneration. This could help provide a scientific basis for the restoration and protection of forest ecosystems. The Liangshui Nature Reserve of Yichun City, Heilongjiang Province, is a coniferous and broadleaved mixed forest. In this study, we assess the regeneration probability of coniferous saplings (CRP) in natural forests in 665 temporary plots in the Liangshui Nature Reserve. Using Sentinel-1 and Sentinel-2 images provided by the European Space Agency, as well as digital elevation model (DEM) data, we calculated the vegetation index, microwave vegetation index (RVI S1), VV, VH, texture features, slope, and DEM and combined them with field survey data to construct a logistic regression (LR) model, geographically weighted logistic regression (GWLR) model, random forest (RF) model, and multilayer perceptron (MLP) model to predict and analyze the CRP value of each pixel in the study area. The accuracy of the models was evaluated with the average values of the area under the ROC curve (AUC), kappa coefficient (KAPPA), root mean square error (RMSE), and mean absolute error (MAE) verified by five-fold cross-validation. The results showed that the RF model had the highest accuracy. The variable factor with the greatest impact on CRP was the DEM. The construction of the GWLR model considered more spatial factors and had a lower residual Moran index value. The four models had higher CRP prediction results in the low-latitude and low-longitude regions of the study area, and in the high-latitude and high-longitude regions of the study area, most pixels had a CRP value of 0 (i.e., no coniferous sapling regeneration occurred).

Keywords: regeneration probability; Sentinel-1 and Sentinel-2 images; logistic regression (LR) model; geographically weighted logistic regression (GWLR) model; random forest (RF) model; multilayer perceptron (MLP) model

Citation: Zhao, H.; Sun, Y.; Jia, W.; Wang, F.; Zhao, Z.; Wu, S. Study on the Regeneration Probability of Understory Coniferous Saplings in the Liangshui Nature Reserve Based on Four Modeling Techniques. *Remote Sens.* **2023**, *15*, 4869. https://doi.org/10.3390/rs15194869

Academic Editors: Iñigo Molina, Jan Komarek and Marlena Kycko

Received: 22 August 2023
Revised: 4 October 2023
Accepted: 6 October 2023
Published: 8 October 2023

1. Introduction

Forests represent an important resource for human beings [1] and have great social, economic, and ecological value [2–4]. With the continuous development and progress of society, people's awareness of protecting their ecological environment is constantly increasing. An understanding of the status of forest regeneration is essential for protecting the ecological environment [5]. Forest regeneration is an important ecological process and has always been one of the main fields of research related to dynamic changes in forest ecosystems [6]. The status of forest regeneration is mainly determined by the composition and structure of

saplings, which determine the future structure of a forest. Therefore, a study of the laws of forest regeneration via saplings can effectively capture forest ecosystem dynamics. In addition, good regeneration plays a crucial role in the restoration of secondary forests and is one of the prerequisites for ensuring sustainable forest management.

Sapling regeneration refers to the use of saplings in a forest to supplement or replace existing old trees, which can maintain the health and stability of the forest [7]. Some scholars have conducted detailed research on sapling regeneration. Angela E. Boag et al. [8] used a logistic regression (LR) model and a generalized linear mixed effects model to determine the factors driving sapling regeneration. Feng Liu et al. [9] analyzed the impact of microenvironmental variables on the biomass accumulation of saplings in forests by using a nonlinear mixed model. Other scholars have explored the factors influencing sapling regeneration and concluded the following: human activities affect the regeneration of understory saplings in pine oak forests [7]; expanding gap afforestation (i.e., reducing forest density) is beneficial for the regeneration of saplings in secondary forests [7,9]; soil temperature and soil permeability affect understory regeneration in secondary forests [9]; and stand type influences biomass accumulation in oak secondary forests [10]. Few scholars have analyzed a specific tree species when studying the law of sapling regeneration. We believe this may have led to inaccuracies because the sapling regeneration law of different tree species cannot be generalized. There has been very little research on the regeneration of coniferous saplings. Many scholars have studied only the law of sapling regeneration without making detailed predictions about the probability of sapling regeneration at each specific location in the study area. In addition, most scholars have conducted little research on combining remote sensing variables in analyzing the regeneration patterns of saplings. The main purpose of this study was to use understory coniferous saplings as the research object and combine the measured data of the sample plots with remote sensing variables extracted from Sentinel-1, Sentinel-2, and DEM to construct four models to predict the CRP size at each specific location in the study area. We also analyzed the factors that had the greatest impact on the regeneration of coniferous saplings based on the four constructed models.

Remote sensing has simplified data acquisition [11]. Some scholars have used remote sensing methods to obtain data for analyzing vegetation [12–14]. Compared to traditional methods of obtaining data, remote sensing has the following advantages: the efficiency of obtaining data has greatly increased [15]; more diversified data sources [16] can be used to more fully describe forest stand characteristics [17]; and more variable factors can be obtained to improve the accuracy and interpretability of models [15]. In this study, we combined field survey data with Sentinel-1 and Sentinel-2 images provided by the European Space Agency, as well as DEM data, to calculate the vegetation index, texture features, slope, and DEM variables to analyze the size and spatial distribution of the CRP in the study area.

In previous studies, most scholars used the method of constructing models to analyze and study the spatial distribution and factors influencing CRP [8,9]. The LR model is a binary classification model that can be used to predict the probability of positive samples with good accuracy [18]. The LR model has been widely used for predicting the probability of events and has also been widely applied in forestry. The premise for constructing the LR model is to assume the stability of space, but in forestry research, this is an ideal state, and the environmental factors in different spatial locations vary [19]. This suggests that the factors affecting the regeneration of saplings in different geographical locations in a forest may differ. In this case, the accuracy of the LR model for CRP prediction would not be sufficient. Therefore, some scholars have proposed the geographically weighted regression (GWR) model, incorporating a spatial variation function [20,21]. The GWR model can be used effectively for spatial nonstationary analysis in dynamic environments and has been widely used in multiple fields [22–24]. However, LR models and GWR models need to be constructed after screening variables, and forest regeneration is an extremely complex ecological process that is constantly influenced by various factors. Clearly, based on these

selected variables, an accurate analysis of understory sapling regeneration is challenging. As shown in previous research, the implementation of machine learning algorithms does not limit the number of variables [25,26], and machine learning has many advantages over traditional models [26]: (I) there are no prerequisite requirements for data types and formats in machine learning and (II) machine learning can be used to effectively handle the complex relationships between independent and dependent variables and to delve deeper into the connections between data. Two types of machine learning algorithms, RF and MLP, have been widely used [27–32]. RF is a classification or regression model [33] built based on decision trees, and the MLP model is based on the construction of multilayer hidden layers and the mapping output of its results through the activation function.

The Liangshui National Nature Reserve was the study area. We established a relationship model between the CRP and forest variables and analyzed the spatial distribution of the CRP. The basic outline of this study was the following: (I) acquire field survey data and remote sensing data; (II) analyze the distribution patterns of the CRP in different forest types and different latitudes and longitudes; (III) build RF, MLP, GWLR, and LR models; (IV) determine the spatial autocorrelation of model residuals; and (V) utilize the four constructed models for predicting and analyzing the CRP size of each pixel in the Liangshui Nature Reserve.

2. Materials and Methods

2.1. Overview of the Study Area

The study area is in Dailing District, Yichun City, Heilongjiang Province, with geographical coordinates of 128°47′8″ to 128°57′19″E and 47°6′49″ to 47°16′10″N. The zonal vegetation in the study area is a temperate coniferous broadleaved mixed forest mainly composed of Korean pine. The vegetation in the whole area can be divided into 11 formations and 19 associations of 6 vegetation types, including cold temperate coniferous forest, temperate mixed coniferous and broadleaved forest, deciduous broadleaved forest, shrub, meadow, and swamp. The study area is mainly composed of natural forests and the forest structure is predominantly multilayered forests. In addition, the study area belongs to a low mountainous and hilly zone, with rounded mountain tops and asymmetrical slopes on both sides of the mountains. Generally, the southern slopes are short and steep, while the northern slopes are gentle and long. The average slope gradient ranges from 10 to 15 degrees, with some areas having steep slopes of over 20 to 40 degrees. The geographical location of the study area is shown in Figure 1a. Figure 1b presents the subcompartment boundaries (indicated by yellow lines) and the specific distribution of 665 sample points (indicated by blue dots) during the field forest resource survey of Liangshui Nature Reserve in Yichun City, Heilongjiang Province, from May to September 2022, within the subcompartments in the study area, using a preprocessed Sentinel-2 remote sensing image as the base map.

2.2. Data Acquisition

2.2.1. Ground Standard Land Survey

In the Liangshui Nature Reserve of Yichun City, Heilongjiang Province, from May to September 2022, we surveyed a total of 665 sample plots. We used a random sampling method to determine the sample plots [34], and the area of each sample plot was 0.06 ha [34]. After the sample plot was identified, the stand type, diameter breast height (DBH, cm), and maximum tree height (MTH, m) were recorded. The single wood volume was calculated based on the one-dimensional volume table of the Liangshui Nature Reserve (the volume table formula is shown in Table 1). At the same time, we conducted a survey of the CRP under the forest in these 665 sample plots. For this survey, coniferous saplings were defined as coniferous tree species with a height greater than 30 cm [34] and a DBH less than 5 cm [34]. In addition, measurements were required for coniferous saplings of sample plots, including the height (H, cm), basal diameter (BD, cm), age, and DBH for saplings with a height greater than 130 cm [34] (the basic statistics of the saplings are shown in Table 2).

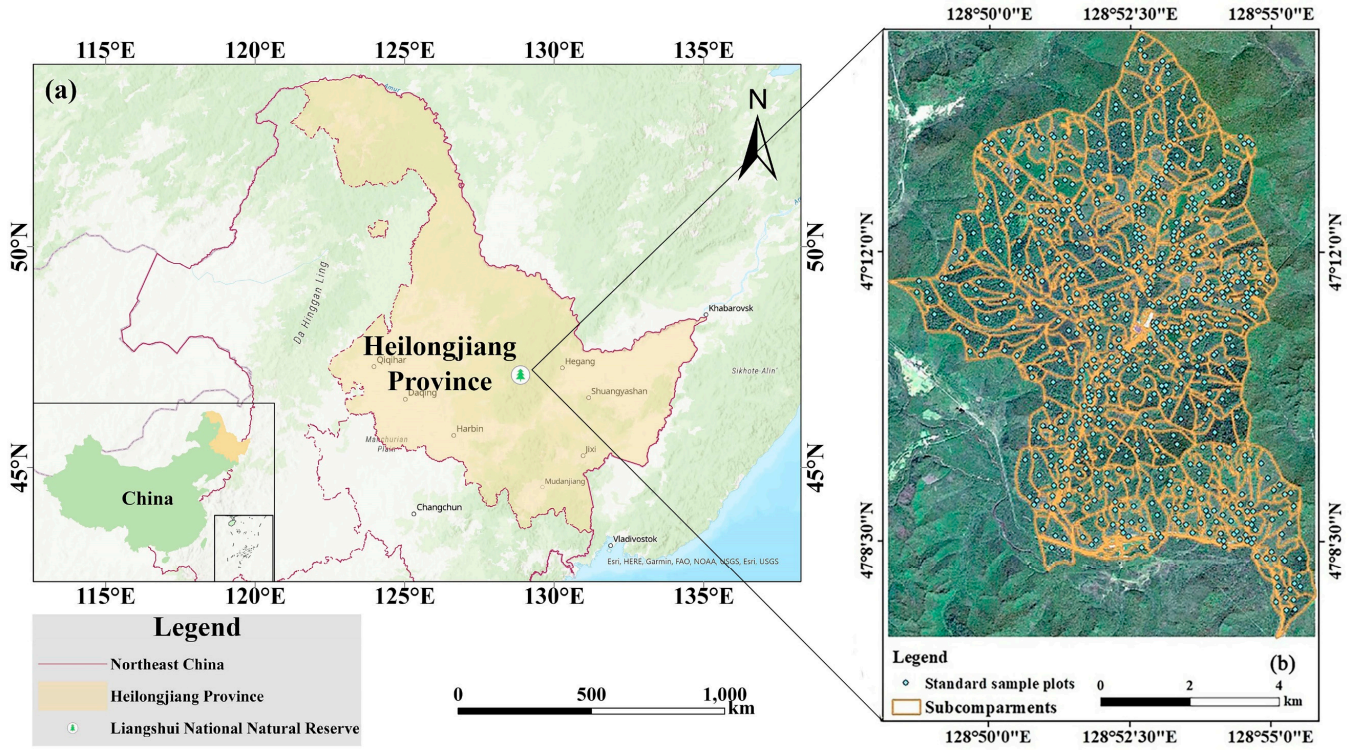

Figure 1. Distribution diagram of the study area and sample plots. (**a**) The specific location of the Liangshui Nature Reserve in China and (**b**) the distribution of ground sample plots in subcompartments.

Table 1. Formulas for tree species volumes.

Tree Species Type	Volume Formula
Korean pine	0.00010339412 × (−0.005162178 + 0.975389083 × DBH)^ (2.5550714)
Spruce	0.000097559294 × (−0.023269474 + 0.979033877 × DBH)^ (2.6082001)
Fir	0.00012553802 × (−0.14050637 + 0.976669654 × DBH)^ (2.5301655)
Camphor pine	0.0002380777 × (−0.1661345 + 0.983825482 × DBH)^ (2.3888099)
Larch	0.00005016824 × (−0.1661345 + 0.983825482 × DBH)^1.7582894 × (1.6504613 + 0.78031609 × (−0.1661345 + 0.983825482 × DBH) −0.0076188678 × (−0.1661345 + 0.983825482×DBH)^2)^ (1.1496653)
Pinus densiflora	0.00016773252 × (0.1539054215 + 0.981705489 × DBH)^ (2.2855543)
Fraxinus mandshurica Rupr.	0.000041960698 × (−0.0283700973 + 0.969811198 ×DBH)^ (1.9094595) × (5.6382753 + 0.64085 × (−0.0283700973 + 0.969811198 × DBH) −0.0056371339 × (−0.0283700973 + 0.969811198 ×DBH)^2)^ (1.0413892)
Juglans mandshurica	0.000041960698 × (−0.1068104174 + 0.975403018×DBH)^ (1.9094595) × (6.5706028 + 0.51071923 × (−0.1068104174 + 0.975403018 × DBH) −0.0034904923 × (−0.1068104174 + 0.975403018 × DBH)^2)^ (1.0413892)
Phellodendron	0.00018200258 × (−0.2516967596 + 0.972900665 × DBH)^ (2.3187749)
Linden tree	0.000041960698 × (0.2250730369 + 0.964592149 × DBH)^ (1.9094595) × (5.2592429 + 0.5670384 × (0.2250730369 + 0.964592149 × DBH) −0.0038177352 × (0.2250730369 + 0.964592149 × DBH)^2)^ (1.0413892)
Oak	0.00025462482 × (0.1751205585 + 0.986711062 × DBH)^ (2.1935242)
Elm	0.00013344177 × (−0.120162996 + 0.971592141 × DBH)^ (2.4489629)

Table 1. *Cont.*

Tree Species Type	Volume Formula
Maple birch	0.000041960698 × (0.040314124 + 0.957532468 × DBH)^ (1.9094595) × (7.0086039 + 0.6791334 × (0.040314124 + 0.957532468 × DBH) −0.0063965703 × (0.040314124 + 0.957532468 × DBH)^ (2))^ (1.0413892)
Black birch	0.000052786451 × (−0.4899312906 + 0.995171441 × DBH)^ (1.7947313) × (6.2804214 + 0.46824315 × (−0.4899312906 + 0.995171441 × DBH) −0.0046635886 × (−0.4899312906 + 0.995171441 × DBH)^2)^ (1.0712623)

Table 2. The basic statistics of coniferous saplings.

Sapling Height		Min	SD	Mean	Max
<130 cm	Basal diameter (BD, cm)	0.5	0.645	1.476	3.963
	Diameter breast height (DBH, cm)	–	–	–	–
	Age	5	2.796	9.411	16
	Height (H, cm)	5	28.204	84.263	129
≥130 cm	Basal diameter (BD, cm)	0.7	1.522	3.635	7.512
	Diameter breast height (DBH, cm)	0.1	1.080	0.532	2.360
	Age	10	2.274	14.115	17
	Height (H, cm)	132	97.831	286.825	610

2.2.2. Remote Sensing Data Acquisition

This study obtained field survey data during the summer of 2022 from May to September. Based on the timing of the field survey data, we acquired Sentinel-2 imagery from 11 June 2022, which was of good quality during the summer. As there were no Sentinel-1 data available for the entire year of 2022, we selected Sentinel-1 imagery from 15 July 2021, which was also of good quality during the summer. Additionally, we utilized the ASTER GDEM 30 m resolution digital elevation dataset. We set each sample plot area to 0.06 ha (approximately 25 m× 25 m), and we uniformly resampled the Sentinel-2, Sentinel-1, and DEM images to a spatial resolution of 25 m.

Sentinel-1 is an Earth observation satellite in the Copernicus Programme (i.e., GMES) of the European Space Agency. It is composed of two satellites and carries C band synthetic aperture radar, which can provide continuous images (day, night, and various weather) [35]. While the 5.5 cm wavelength of Sentinel-1 cannot penetrate a forest canopy to analyze understory vegetation, the texture features and microwave remote sensing index variables extracted from Sentinel-1 in this study could to some extent characterize the forest stand characteristics. The GRD data of the Sentinel-1 remote sensing image were used in this study, and the main preprocessing was completed using SNAP 9.0.0 software provided by the European Space Agency. The operation steps included thermal noise removal, orbit file correction, speckle filtering, radiometric calibration, data format conversion, and Doppler terrain correction. Finally, we obtained Sentinel-1 remote sensing images using VV and VH polarization methods (where VV polarization represents vertical transmission data and vertical reception data, and VH polarization represents vertical transmission data and horizontal reception data) [36,37], and we calculated the radar vegetation index (RVI). Because speckle filtering and terrain correction can damage the texture of images [38,39], we extracted texture features from images without performing speckle filtering and terrain correction. Finally, 16 texture feature variables were output (as shown in Table 3).

The Sentinel-2A satellite is the second satellite of the Global Environment and Security Monitoring program [40]. Sentinel-2 data are the only data with three bands in the red edge range, which makes them very useful for monitoring vegetation health information [41]. The multiple bands provided by Sentinel-2 can be used to effectively describe the site conditions of a forest through calculations, helping to consider the regeneration patterns and factors influencing understory coniferous saplings from multiple perspectives [42]. In this

study, we used Sentinel-2 S1C-level remote sensing images provided by the European Space Agency. The preprocessing steps were the following: downloading Sentinel-2 S1C-level remote sensing images for the corresponding month of the year; using the Sen2Cor atmospheric correction processor based on the radiative transfer model to perform atmospheric correction on multispectral images, obtaining atmospheric bottom reflectance products, and completing the conversion of S1C-level data to S2A-level data; band resampling, consisting of opening the processed S2A-level image in SNAP, resampling all bands to 25 m resolution, and storing the results in ENVI format; and extracting the vegetation index from the Sentinel-2 image data and used ENVI 5.3 software to perform band operations using the Band Math tool. The 18 S-2 vegetation indices are shown in Table 3.

DEM are important data for studying and analyzing terrain, watersheds, and feature recognition. Due to the ability of DEM data to reflect local terrain features at a certain resolution, a large amount of surface morphological information can be extracted from DEM. Based on DEM, features such as slope, aspect, and contour lines can be further calculated. In this study, we calculated the slope and DEM information of the study area based on the downloaded DEM data of the study area (Table 3).

Table 3. Extracted remote sensing factors and terrain factors.

	Vegetation Index	Abbreviation	Calculation Formula
S2-VI	Ratio VI	RVI (S2)	$B8/B4$ [43]
	Difference VI	DVI	$B8–B4$
	Weighted Difference VI	WDVI	$B8–0.5 \times B4$
	Infrared Percentage VI	IPVI	$B8/(B8 + B4)$ [44]
	Perpendicular VI	PVI	$\sin(45°) \times B8–\cos(45°) \times B4$
	Normalized Difference VI	NDVI	$(B8–B4)/(B8 + B4)$
	Transformed Normalized Difference VI	TNDVI	$[(B8–B4)/(B8 + B4) + 0.5]1/2$
	Soil-Adjusted VI	SAVI	$1.5 \times (B8–B4)/8 \times (B8 + B4 + 0.5)$
	Modified Soil-Adjusted VI	MSAVI	$(2–\text{NDVI} \times \text{WDVI}) \times (B8–B4)/8 \times (B8 + B4 + 1–\text{NDVI} \times \text{WDVI})$
	Modified Soil-Adjusted VI2	MSAVI2	$0.5 \times (2 \times (B8 + 1)) –\text{sqrt}\,[(2 \times B8 + 1) \times (2 \times B8 + 1) –8 \times (B8–B4)]$
	Atmospheric Ratio VI	ARVI	$[B8–(2 \times B4–B2)]/[B8 + (2 \times B4–B2)]$
	Normalized Difference Water Index	NDWI	$(B3–B8)/(B3 + B8)$
	Normalized Difference Built-up Index	NDBI	$(B11–B8)/(B11 + B8)$
	Green Atmospherically Resistant Index	GARI	$(B8–(B3–1.7 \times (B2–B4)))/(B8 + (B3–1.7 \times (B2–B4)))$
	Optimized Soil-Adjusted VI	OSAVI	$1.5 \times (B8–B4)/(B8 + B4 + 0.16)$
	VI Green	VIG	$(B3–B4)/(B3 + B4)$
	Normalized Difference Moisture Index	NDMI	$(B8–B11)/(B8 + B11)$
	Normalized Difference Senescent VI	NDSVI	$(B11–B4)/(B11 + B4)$
S1-Textural	Mean	VH_MEA VV_MEA	$ME = \sum_{i,j=0}^{N-1} i * P_{i,j}$
	Variance	VH_VAR VV_VAR	$VA = \sum_{i,j=0}^{N-1} i * P_{i,j}(i - ME)^2$
	Homogeneity	VH_HOM VV_HOM	$HO = \sum_{i,j=0}^{N-1} i * \frac{P_{i,j}}{1+(i-j)^2}$
	Contrast	VH_CON VV_CON	$CO = \sum_{i,j=0}^{N-1} i * P_{i,j}(i - j)^2$
	Dissimilarity	VH_DIS VV_DIS	$DI = \sum_{i,j=0}^{N-1} i * P_{i,j}\lvert i - j\rvert$
	Entropy	VH_ENT VV_ENT	$EN = \sum_{i,j=0}^{N-1} i * P_{i,j}(-\ln p_{i.j})$
	Second Moment	VH_ASM VV_ASM	$SM = \sum_{i,j=0}^{N-1} i * P_{i,j}$
	Correlation	VH_COR VV_COR	$CR = \sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-ME)*(j-ME)}{\sqrt{VA_i * VA_j}}\right]$

Table 3. *Cont.*

	Vegetation Index	Abbreviation	Calculation Formula
S1	–	VV, VH	–
	Radar VI	RVI (S1)	VH/VV
DEM	DEM (m)	–	Composed of elevation values of points on the ground
	Slope (°)	–	Rate of elevation change at a point on the ground

B2: blue band; ***B3***: green band; ***B4***: red band; ***B8***: near-infrared band (wide); ***B11***: shortwave infrared band; ***i***: the gray level is an *i* pixel value; $p_{i,j}$: the probability that a pixel with a grayscale value of *i* is at a certain distance from another pixel with a grayscale value of *j*.

The main content of the technical route is shown in Figure 2, which can be divided into 4 parts: extraction information, model construction, model evaluation, and spatial distribution of the CRP.

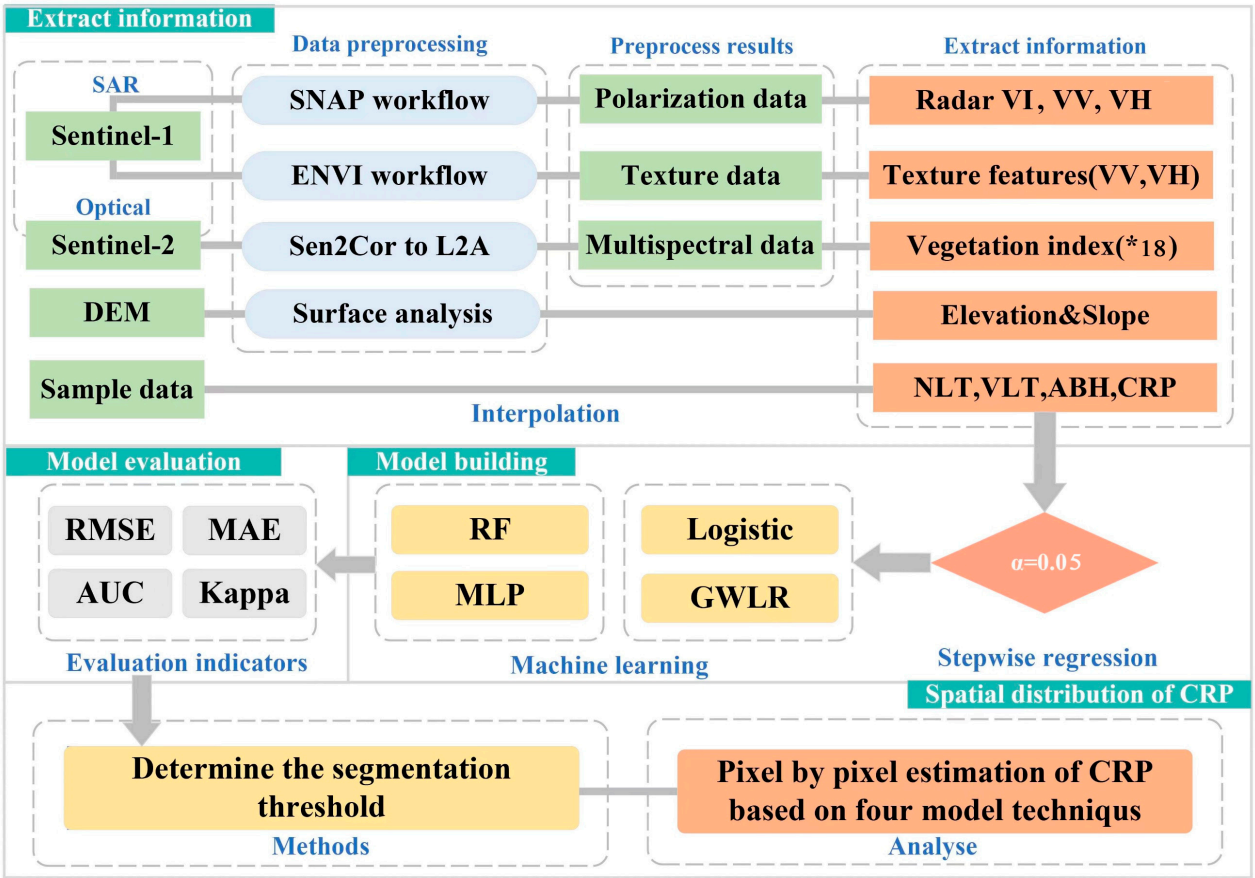

Figure 2. Technical route. "*18" represents that we extracted 18 Sentinel 2 vegetation indices.

2.2.3. Variable Screening

Four stand factors were calculated based on data from 665 sample plots (the basic statistics are shown in Table 4), including the average DBH (AD, cm), maximum heights of the tree (MTH, m), volume of living trees per hectare (VLT, m^3/ha), and number of living trees per hectare (NLT, n/ha). Thirty-seven remote sensing factors, 2 terrain factors, and data from the sample plots were used as the independent variables to construct the RF and MLP models with CRP as the dependent variable. In addition, stepwise regression was used to screen significant variable factors (α = 0.05); at the same time, the variable factors that caused multicollinearity were screened and eliminated. Finally, the average DBH (AD,

cm), volume of living trees per hectare (VLT, m^3/ha), VV_VAR, VH_CON, GARI, and DEM were selected as the modeling factors to construct the LR and GWLR models. Based on the 665 known plots surveyed, the stand types were divided into 6 types: broadleaf mixed forest (BMF), broadleaf relatively pure forest (BRPF), coniferous-broadleaved mixed forest (CBMF), coniferous pure forest (CPF), coniferous mixed forest (CMF), and coniferous relatively pure forest (CRPF). The number and percentages of sample plots for the stand types in the 665 sample plots are shown in Table 5.

Table 4. Basic Statistics of measured data.

Variable	Min	SD	Mean	Max
CRP	0	0.431	0.755	1
NLT (n/ha)	200	356.158	805.590	3083.333
AD (cm)	9.85	4.210	19.653	40.85
VLT (m^3/ha)	31.081	59.34893	193.807	402.243
MTH (m)	8.8	4.899	20.913	39.117

Table 5. Statistics on the number and percentages of sample plots of different stand types.

Stand Type	Number of Sample Plots	Percentage
Broadleaf Mixed Forest (BMF)	106	15.9%
Broadleaf Relatively Pure Forest (BRPF)	20	3.01%
Coniferous Broadleaved Mixed Forest (CBMF)	244	36.7%
Coniferous Pure Forest (CPF)	26	3.91%
Coniferous Mixed Forest (CMF)	152	22.9%
Coniferous Relatively Pure Forest (CRPF)	117	17.6%

To further assess whether there was spatial heterogeneity in the CRP values of our 665 sample plots, we conducted statistics on the mean CRP values based on the stand type and latitude and longitude directions (as shown in Tables 6 and 7) and used Origin 2022 software to plot the statistical results (as shown in Figure 3). The distribution of the CRP along the latitude and longitude lines exhibited spatial heterogeneity in both the same and different stands types.

Table 6. CRP statistics in the latitudinal direction of different stand types.

47° & Stand Type	7′	8′	9′	10′	11′	12′	13′	14′	CRP Mean
Broadleaf Mixed Forest (BMF)	1	0.724	0.478	1	0.75	0.727	0.7	0	0.672
Broadleaf Relatively Pure Forest (BRPF)	0	0.833	1	1	0.75	0.625	1	0	0.651
Coniferous Broadleaved Mixed Forest (CBMF)	0.5	0.794	0.759	0.759	0.696	0.725	0.643	1	0.735
Coniferous Pure Forest (CPF)	0	1	0.75	0.75	1	0.462	1	0	0.620
Coniferous Mixed Forest (CMF)	1	0.909	0.9	0.941	0.911	0.909	0.542	0	0.764
Coniferous Relatively Pure Forest (CRPF)	1	0.833	0.905	0.895	0.794	0.565	0.467	0.4	0.732

Table 7. CRP statistics in the longitudinal direction of different stand types.

128° & Stand Type	48′	49′	50′	51′	52′	53′	54′	55′	CRP Mean
Broadleaf Mixed Forest (BMF)	0	1	0.6	0.905	0.769	0.5	0.636	0.750	0.645
Broadleaf Relatively Pure Forest (BRPF)	0	0	0	0.8	0.667	0.818	0.800	0	0.386

Table 7. *Cont.*

Stand Type \ 128° &	48′	49′	50′	51′	52′	53′	54′	55′	CRP Mean
Coniferous Broadleaved Mixed Forest (CBMF)	1	0.444	0.5	0.929	0.868	0.686	0.541	0.833	0.725
Coniferous Pure Forest (CPF)	0	1	0.667	0.667	0.8	0.714	1	0	0.606
Coniferous Mixed Forest (CMF)	0	0.625	0.889	0.944	0.897	0.72	0.636	1	0.714
Coniferous Relatively Pure Forest (CRPF)	1	0.714	0.583	0.833	0.786	0.667	0.739	0.667	0.749

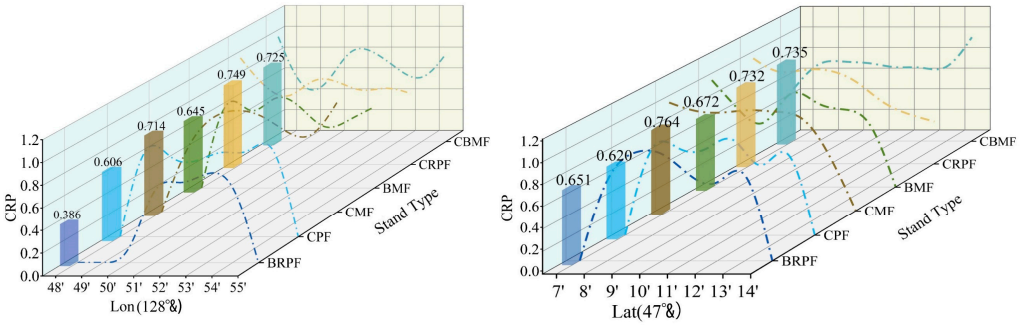

Figure 3. CRP distribution rule analysis. The curve shows the change trend of the CRP at different longitudes and latitudes under the same stand type. The bar chart shows the average CRP distribution along the longitudinal and latitudinal directions under the same stand type.

2.2.4. Study on CRP based on the LR Model

The LR model is simple to calculate, easy to interpret and understand, and performs well on linearly separable or approximately separable datasets. In many practical cases, the response variable is basically binary. That is, there are two possible results: 0 (not occurring) or 1 (indeed occurring). Here, we set CRP $\neq$ 0 (y = 1, which means there is a coniferous sapling regeneration occurring in the sample plot) as P, and CRP = 0 (y = 0, meaning there is no coniferous sapling regeneration occurring at the sample plot) was set to 1-P [45]. Logistic regression was established between the occurrence probability of coniferous sapling regeneration and its respective variables via the Python programming language. The construction of the LR model was based on Python's "sklearn.linear_model" library. Our parameter selection was based on a grid search. The grid search method was based on Python's "sklearn.model_selection" library. The mathematical expression is shown in Formula (1):

$$Logit(P) = ln(P/(1-P)) = \beta_0 + \sum_{k=1}^{5} \beta_k x_{ik} \quad (1)$$

Here, P represents the probability of coniferous sapling regeneration occurring, β_0~β_k represents the regression coefficient of the model, and k is the number of independent variables.

2.2.5. Study on CRP Based on the GWLR Model

GWLR is an extension of LR that can consider the weight of geographical factors in space. It can capture spatial heterogeneity and local nonlinear relationships, thus improving the predictive ability of the model. The phenomenon of changes in the relationship or structure between variables due to differences in geographical location is called spatial nonstationarity. The GWLR model is an extension of the LR model, but it accounts for spatial location factors based on LR and uses weighting to estimate parameters for each coordinate point. The estimation of GWLR model parameters is local rather than global, and each position has corresponding parameter estimation coefficients [19]. Geographically weighted regression was established between the occurrence probability of coniferous sapling regeneration and its respective variables using the Python programming language.

The construction of the GWLR model was based on Python's "mgwr.gwr" library. First, we used Python's "mgwr.sel_bw" library to select the optimal bandwidth based on the coordinates of the sample location. Then, we selected the model construction type in Python's "spglm.family" library. Because we were studying binary classification problems, we chose the "binomial" parameter to be passed into the GWR model. The mathematical expression is shown in Formula (2):

$$Logit(P(u_i, v_i)) = ln(P(u_i, v_i)/(1 - P(u_i, v_i))) = \beta_0(u_i, v_i) + \sum_{k=1}^{5} \beta_k(u_i, v_i)x_{ik} \quad (2)$$

where $\beta_0(u_i, v_i) \sim \beta_k(u_i, v_i)$ represents the coefficient of the GWLR model at position i.

2.2.6. Study on CRP Based on the RF Model

Random forest is an ensemble learning method that combines multiple decision trees, with high prediction accuracy and robustness. It can handle complex relationships between high-dimensional data and features, while also possessing a certain degree of noise resistance. RF modeling is usually used for classification and can also be used for regression [46]. An RF model is constructed by assembling multiple regression trees, training the training samples, and finally selecting the optimal solution by averaging (voting on) the decision tree results. The greatest advantage of the RF model is that it has no restrictions on the dataset. In addition, the RF model is based on many decision trees in the construction process, avoiding overfitting to a great extent.

Figure 4 shows the principle of the RF model developed with the Python programming language. The construction of the RF model was based on Python's "sklearn.ensemble" library. First, we determined the random forest hyperparameters. The determination of random forest hyperparameters was achieved through a random search and grid search. We used Python's "sklearn.model_selection" library for the random search and grid search. We searched for the optimal hyperparameter combination within the defined range of the random search and grid search. After inputting the training data into the RF model, N bootstrap resamplings were performed on the training data, and a decision tree was constructed for each sample. A total of N decision trees was constructed, and the return value of each decision tree was averaged (voted on) to calculate the output result of the RF model. We used the random search and grid search to find an optimal hyperparameter combination. The hyperparameters of the RF model we constructed included N_Estimators = 220, the number of trees in the random forest, which was set to 220; Max_Depth = 500, the maximum depth of each decision tree in the random forest, which was set to 500; Min_Samples_Leaf = 11, the minimum number of samples required to be at a leaf node in each decision tree, which was set to 11; Min_Samples_Split = 6, the minimum number of samples required to split an internal node in each decision tree, which was set to 6; and Max_Features = "sqrt", the number of features to consider when looking for the best split at each node (here, "sqrt" indicates the square root of the total number of features considered).

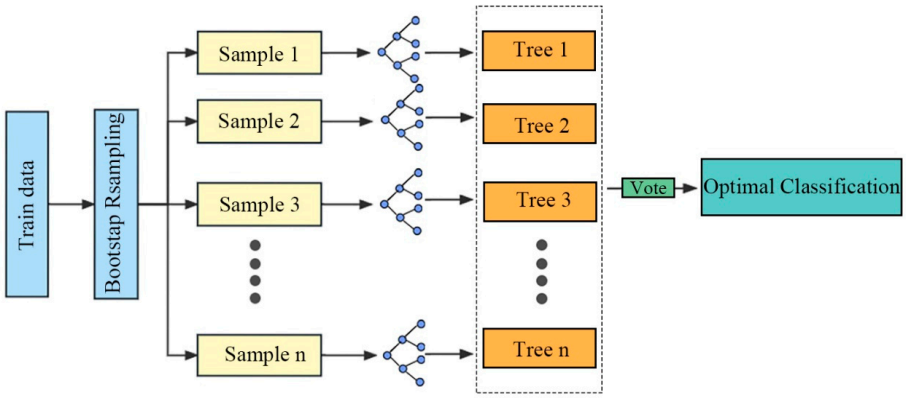

Figure 4. Schematic diagram of the RF model.

2.2.7. Study on CRP Based on the MLP Model

MLP is a highly flexible and expressive model that can handle complex nonlinear relationships and many features. It can improve its predictive performance by adjusting its network structure and parameters. The MLP model (Figure 5) is a common feedforward neural network model that connects input and output layers by one or more hidden layers. Each hidden layer is composed of multiple neurons, which transform the weighted sum of input signals into output signals through an activation function and then transfer them to the next layer. The output of the output layer is determined based on the required tasks [31]. In the process of training the MLP model, we calculated the update direction of the parameters according to the gradient of the loss function through a backpropagation algorithm and transmitted it back to the network. Using random gradient descent to update the parameters, Figure 5 shows the principle of the MLP model using the Python programming language. The construction of the MLP model was based on Python's "tensorflow" library. We built our own input layer, hidden layer, and output layer for the MLP. We defined the activation function as "sigmoid". To prevent overfitting, we used an early stop strategy during the model construction process and set a loss function to monitor the accuracy of the MLP model. After inputting training data into the MLP model, the MLP model iteratively built a model based on the set model parameters, found the lowest "Val Loss" value, and saved the model weights between the input layers and hidden layers, as well as hidden layers and output layers, finally returning an optimal model. The curve on the right side of Figure 5 shows the variation curve of the MAE between the training set and the testing set with the number of iterations (EPOCH) during MLP model fitting. The curve on the right side of Figure 5 shows that when the number of iterations of the model reached 579, the model underwent overfitting, resulting in a significant decrease in the MAE of the training set (a significant increase in the accuracy of the training set) and a gradual increase in the MAE of the testing set (a gradual decrease in the accuracy of the testing set). Therefore, we returned the MLP model weight for the 579th iteration. Next, based on the optimal model, the CRP was predicted and analyzed. The hyperparameters of the MLP model we constructed are layers = 4, i.e., the MLP model had 4 hidden layers; loss = "mae ", i.e., the loss function of the MLP model was "mae"; activation = "sigmoid", i.e., the activation function of the MLP model was "sigmoid"; and learning rate = 0.001, i.e., the learning rate of the MLP model was 0.001.

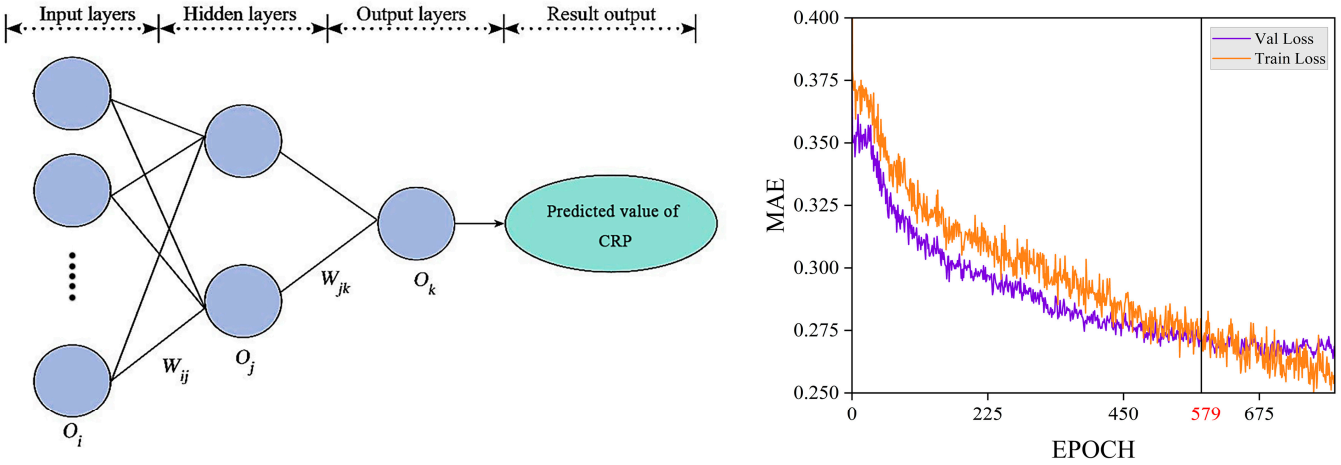

Figure 5. Schematic diagram of the MLP model. O represents the neuron, W represents the weight, and the broken line represents the variation in MAE with the number of iterations (EPOCH). 579: we return the model weights for the 579th iteration.

2.3. Model Evaluation

We obtained the model fitting parameters, evaluated the model prediction accuracy through a 5-fold cross-validation method (Figure 6) using the Python programming lan-

guage, and evaluated the model prediction accuracy using AUC, KAPPA, RMSE, and MAE as indicators. The ratio of our validation set to the training set was 1:4, the number of samples in the training set was 532, and the number of samples in the validation set was 133. The smaller the RMSE and MAE values were, the better the fitting results of the model. Higher KAPPA values represented better classification results for the models. The AUC (defined as the area enclosed by the coordinate axis under the ROC curve, with the false positive rate (FPR) as the abscissa and the true positive rate (TPR) as the ordinate of the ROC curve) is not greater than 1. The range of AUC values is between 0.5 and 1. The closer the AUC is to 1.0, the higher the authenticity of the detection method. When the AUC is equal to 0.5, the authenticity is lower and has no practical value.

$$RMSE = \frac{1}{k}\sum_{j=1}^{k} RMSE_j = \frac{1}{k}\sum_{j=1}^{k}\sqrt{\frac{1}{n_j}\sum_{i=1}^{n_j}(O_{ij}-P_{ij})^2} \tag{3}$$

$$MAE = \frac{1}{k}\sum_{j=1}^{k} MAE_j = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}|O_{ij}-P_{ij}|\right) \tag{4}$$

$$KAPPA = \frac{\frac{\sum_{i=1}^{c} T_i}{n} - \frac{\sum_{i=1}^{c} a_i * b_i}{n^2}}{1 - \frac{\sum_{i=1}^{c} a_i * b_i}{n^2}} \tag{5}$$

$$FPR = \frac{TP}{(TP+FN)} \tag{6}$$

$$TPR = \frac{FP}{(FP+TN)} \tag{7}$$

where k is the number of cross-validation times, i.e., $k = N$; O_{ij}, P_{ij} represent the i-th observation value of the j-th time and the predicted value of the model n_j represents the number of samples for the j-th time; $RMSE_j$, MAE_j represent the j-th root mean square error and the mean absolute deviation, respectively; C is the total number of categories, T_i is the number of samples correctly classified for each category, a_i is the actual number of samples for each category, b_i is the number of samples for each class predicted, and n is the total number of samples; TP represents the number of true-positive instances, which are the positive classes correctly predicted as positive; FN represents the number of false-negative instances, which are the positive classes incorrectly predicted as negative; FP represents the number of false-positive instances, which are the negative classes incorrectly predicted as positive; and TN represents the number of true-negative instances, which are the negative classes correctly predicted as negative.

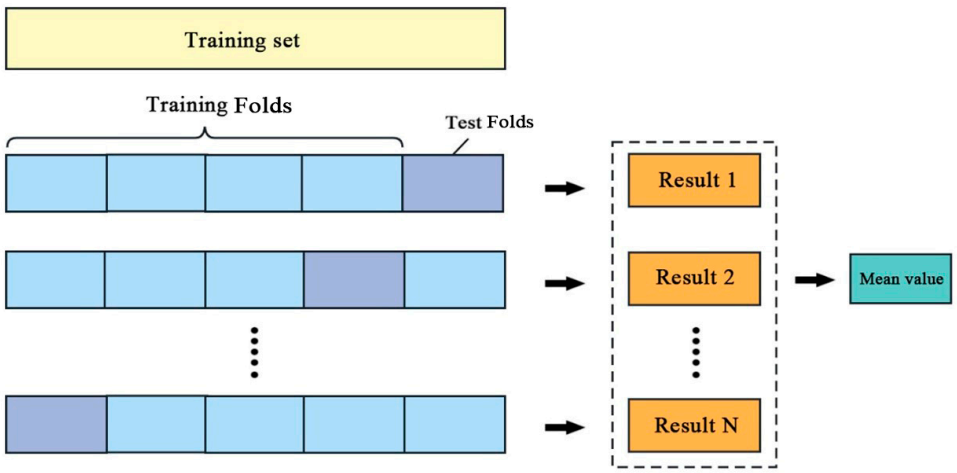

Figure 6. K-fold cross-validation diagram.

2.4. Spatial Autocorrelation Test of Model Residuals

Spatial autocorrelation analysis refers to the distribution, amplitude, and similarity of the same variable in different spatial locations, which mainly measures the degree of aggregation of attribute values of spatial units [47]. Moran's I is the most widely used indicator for measuring spatial autocorrelation in various spatial statistics [48,49], and Moran's I statistic was originally proposed by Moran et al. [50]. The value of Moran's I is generally between –1 and 1; values greater than 0 indicate a positive correlation, a value equal to 0 indicates no correlation, and values less than 0 indicate a negative correlation. The closer Moran's I index value is to 0 for the model residual, the better the performance of the model in reducing spatial autocorrelation, as the index considers spatial location information. We used GeoDa 1.12.1 software to calculate Moran's I, and the method of calculation is shown in Formula (8):

$$I = \frac{z_i - \overline{z}}{\sigma^2} \sum_{j=1, j \neq i}^{n} [w_{ij}(z_j - \overline{z})] \quad (8)$$

In the equation, the average value of variable z is $\overline{z}$; the variance of variable z is σ^2; variable z at sampling points i and j is z_i, z_j (i is not equal to j); and the distance weight between sampling points is denoted as w_{ij}.

3. Results

3.1. Fitting Results of Models

The results of the LR model were statistically significant. As seen in the results in Table 8, the absolute values of the factor coefficients of the six variables based on the LR model can be arranged in the following order: DEM > AD > VH_CON > VLT > VV_VAR > GARI. This indicates that DEM had the greatest impact on CRP, while GARI had the smallest impact on CRP. Specifically, the estimated coefficients of DEM and volume of living trees per hectare (VLT, m^3/ha) were all negative, indicating a negative correlation between these two characteristic variables and the regeneration of understory coniferous saplings throughout the entire study area based on the LR model. The estimated values of the average DBH (AD, cm), VV_VAR, VH_CON, and GARI coefficients were positive, indicating that based on the LR model, the variable was positively correlated with CRP throughout the entire study area.

Table 8. Parameter fitting results of the LR model.

Variable	Estimate	Standard Error	*p* Value	Exp—(Est)
Intercept	1.219	0.094	0.000	3.385
AD (cm)	0.313	0.121	0.010	1.367
VLT (m^3/ha)	−0.180	0.099	0.030	0.835
VV_VAR	0.173	0.106	0.031	1.189
VH_CON	0.311	0.132	0.018	1.365
GARI	0.073	0.099	0.046	1.076
DEM (m)	−0.557	0.107	0.000	0.573

AD (cm): average DBH of forest stands; **VLT (m^3/ha):** standing forest volume per hectare; **DEM (m):** altitude of sample plots; **Estimate:** the estimated coefficient value representing the degree of influence of the variable on the dependent variable; **Standard Error:** the standard error of the estimated coefficient representing the accuracy of the estimation; ***p* Value:** the significance level of the coefficient, indicating whether the variable had a significant impact on the dependent variable; **Exp—(Est):** the exponential value of the coefficient, which is the estimated exponential coefficient, representing the multiple impacts of the variable on the dependent variable.

The GWLR model was constructed after considering a matrix of distance-related spatial weights in the LR model. The variable coefficients of the GWLR model varied with geographical regions (as shown in Table 9), and the average coefficients of the GWLR model were similar to the LR fixed coefficients. However, the coefficients of the GWLR model had a large spatial distribution range. The estimated coefficients of the volume of living

trees per hectare (VLT, m^3/ha), VV_VAR, VH_CON, and GARI feature variables alternated between positive and negative correlations; the DEM and average DBH (AD, cm) variables had a monotonic relationship with CRP in space. The average DBH (AD, cm) variable was positively correlated with the CRP, and the DEM variable was negatively correlated with the CRP. To observe the spatial distribution of GWLR model variable coefficients in the study area, we performed spatial interpolation on the GWLR model variable coefficients using ARCGIS 10.7 software (as shown in Figure 7).

Table 9. Parameter estimations of the GWLR model.

Variable	Min	Lower Quartile	Mean	Median	Upper Quartile	Max
Intercept	0.884	1.162	1.271	1.262	1.386	1.709
AD (cm)	0.032	0.207	0.347	0.287	0.384	0.999
VLT (m^3/ha)	−0.426	−0.308	−0.203	−0.250	−0.103	0.152
VV_VAR	−0.011	0.075	−0.182	0.135	0.257	0.623
VH_CON	−0.104	0.029	0.232	0.209	0.423	0.615
GARI	−0.261	−0.093	0.050	0.055	0.174	0.466
DEM (m)	−2.506	−0.926	−0.819	−0.659	−0.545	−0.381

AD (cm): average DBH of forest stands; **VLT (m^3/ha):** standing forest volume per hectare; **DEM (m):** altitude of sample plots.

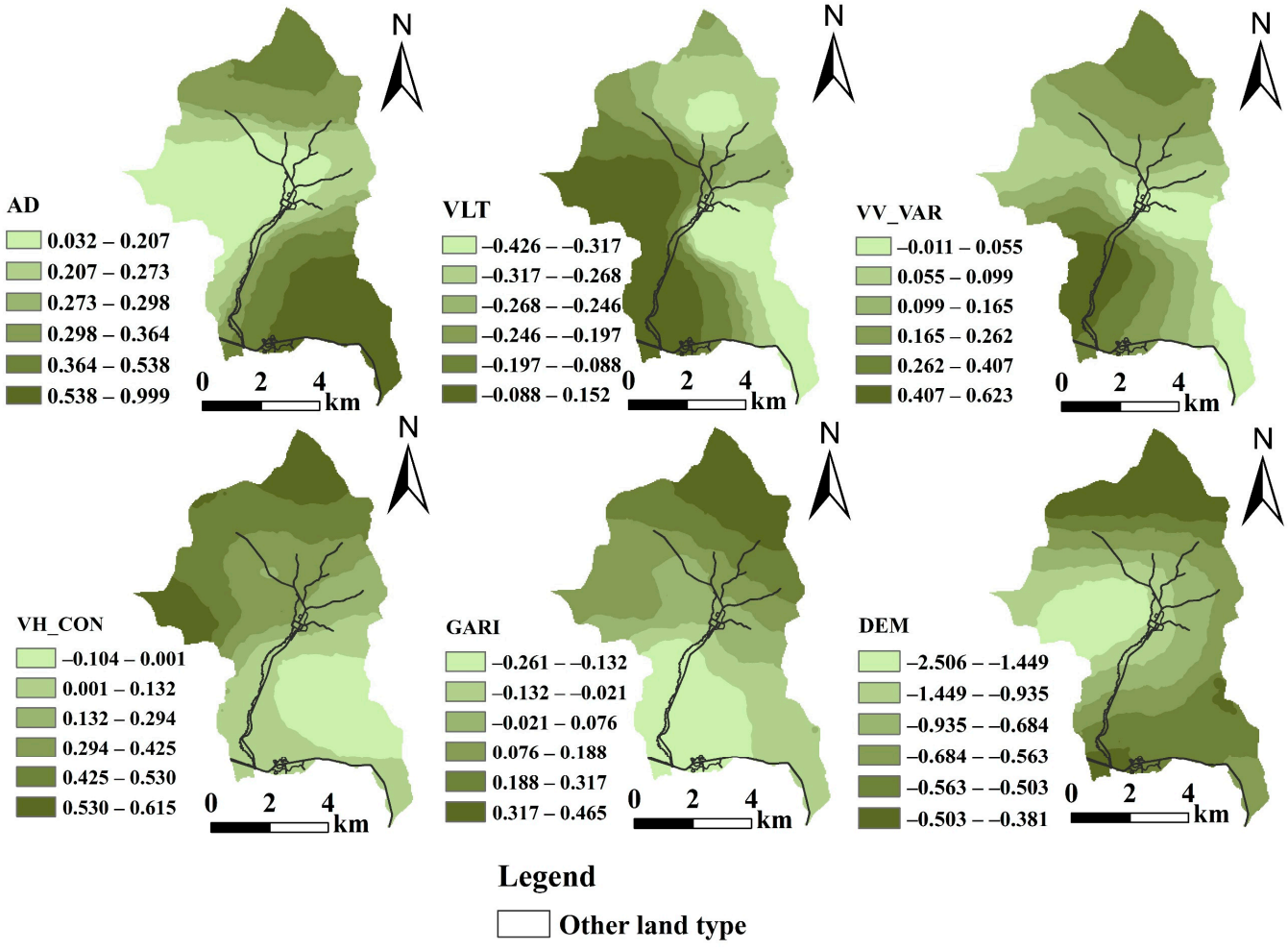

Figure 7. Spatial distribution of GWLR variable coefficients.

Figure 7 shows that the variable average DBH (AD, cm) based on the GWLR model was positively correlated with the CRP throughout the entire study area, and the average DBH (AD, cm) variable coefficient value was greater in high- and low-latitude regions than

in other regions. The variable volume of living trees per hectare (VLT, m^3/ha) was negatively correlated with the CRP in the high-longitude regions and positively or negatively correlated with the CRP in the low-longitude regions. The variable VV_ VAR was positively or negatively correlated with the CRP in mid-latitude and high-longitude regions and positively correlated with the CRP in other regions. The variable VV_ CON was positively or negatively correlated with the CRP in low-latitude regions and positively correlated with the CRP in other regions. The variable GARI was positively correlated with the CRP in the high-latitude regions and negatively correlated with the CRP in the low-latitude regions. The variable DEM had a negative correlation with the CRP throughout the entire study area, and the coefficient's absolute value of the variable DEM in the mid-latitude regions was greater than that in other regions.

3.2. Model Accuracy Evaluation

Table 10 shows the fitting accuracy results of the RF, MLP, LR, and GWLR models after five cross-validation tests. The AUC value of LR was 0.684, indicating that the performance of the model was at a moderate level; KAPPA: 0.225, the KAPPA coefficient was used to evaluate the consistency of the classifier. Compared with the other three models, this value was lower, indicating that the predicted results of the LR model were less consistent with the actual results than the other three models; RMSE, MAE: 0.416, 0.346. Compared with the other three models, the LR model had higher RMSE and MAE values, indicating a larger prediction error of the model than the other three models. The AUC value of the GWLR was 0.751, which compared to the LR model, indicated an improved performance of the GWLR model; KAPPA: 0.277, indicating that the consistency of the GWLR model was relatively good, and the consistency between the predicted results and the actual results improved; RMSE, MAE: 0.400, 0.315, slightly improved compared to the LR model. The AUC value of the MLP was 0.843, so the performance of the MLP model was good, with significant improvement compared to the LR and GWLR models; KAPPA: 0.463, indicating a high consistency between the predicted results of the MLP model and the actual results; RMSE, MAE: 0.350, 0.260, indicating that the MLP model had smaller prediction errors compared to the LR and GWLR models. The AUC value of RF was 0.867, the RF model had the best performance and had a significant improvement compared to other models; KAPPA: 0.561, indicating good consistency of the RF model and high consistency between predicted and actual results; RMSE, MAE: 0.332, 0.240, the RF model had the smallest prediction error compared to other models. Based on the above analysis results, the model fitting accuracy can be sorted as the following: RF > MLP > GWLR > LR. The prediction accuracy of the RF model and MLP model was significantly higher than that of the LR model and GWLR model, fully reflecting the advantages of machine learning in predicting CRP.

Table 10. Model accuracy evaluation.

Model	AUC	Threshold	KAPPA	RMSE	MAE
LR	0.684	0.772	0.225	0.416	0.346
GWLR	0.751	0.811	0.277	0.400	0.315
MLP	0.843	0.677	0.463	0.350	0.260
RF	0.867	0.633	0.561	0.332	0.240

3.3. RF Model Importance Ranking

To further understand the contribution of all variable factors in constructing the RF model, the importance values of the RF model variables were sorted based on the principle of minimum out-of-bag (OOB) error. In Figure 8, the horizontal axis represents the importance score of the RF model during construction, while the vertical axis represents the variable factors. The importance score of DEM feature variables was the highest, and the contribution rate to the establishment of RF models was the largest. We concluded that there was a close relationship between the DEM and CRP.

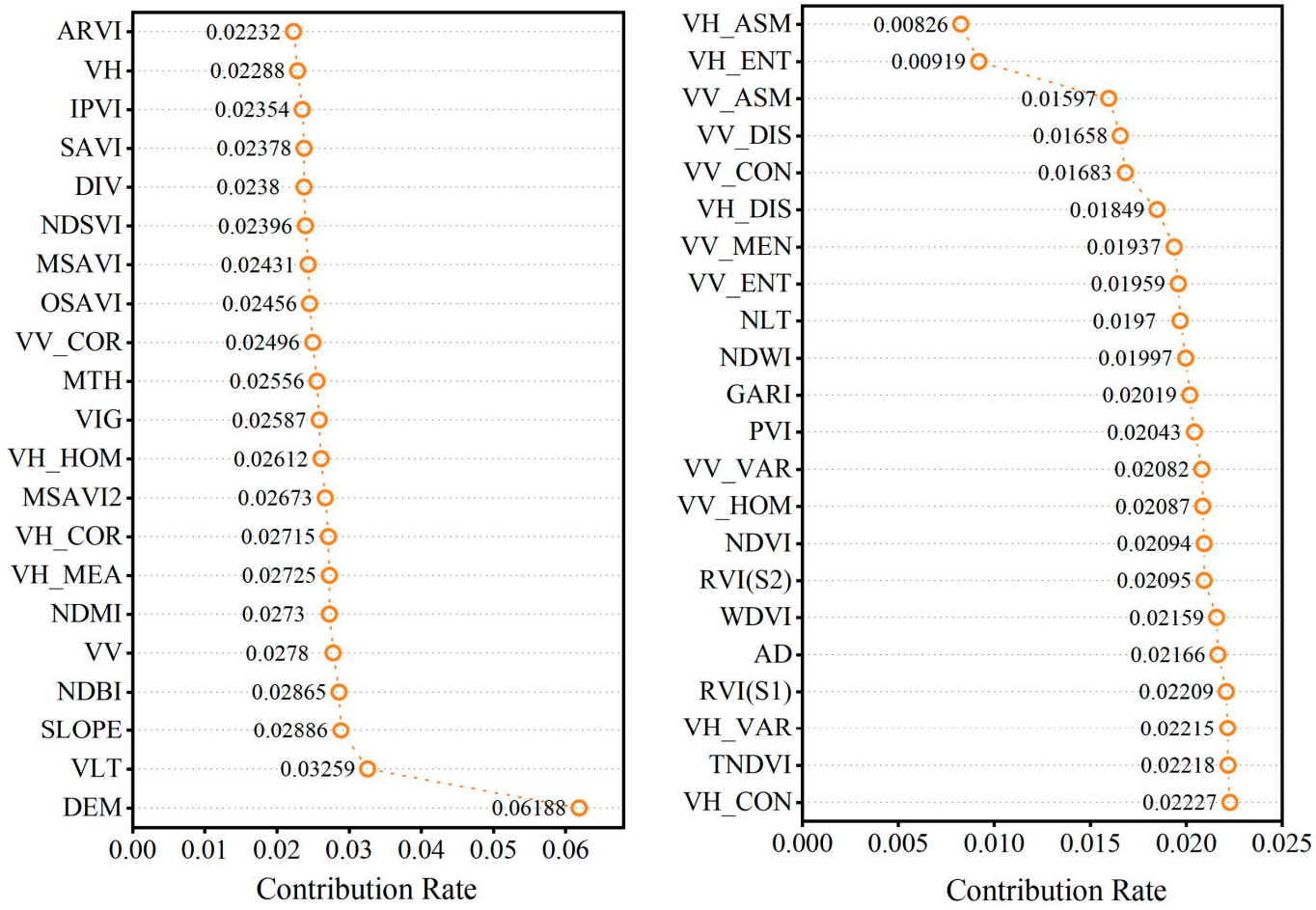

Figure 8. Ranking of importance of variables. Due to the high number of variable factors, they were divided into two columns.

3.4. Analysis of Understory Regeneration Law

We interpolated the forest variable factors with the kriging interpolation method in ARCGIS 10.7 software to obtain the forest variable factor values for each pixel in the entire research area. The CRP size of each pixel in the study area was predicted using four models constructed in the Python programming language. The optimal segmentation threshold for the prediction results of the full sample data model was calculated using the Python programming language. We calculated the optimal segmentation threshold by using the Python programming language to draw the ROC curve. The ROC curve takes the true-positive rate (TPR) as the vertical axis and the false-positive rate (FPR) as the horizontal axis. We calculated the TPR and FPR at different points by changing the threshold. The optimal threshold is usually the point at which the ROC curve is closest to the upper left corner, that is, the point with a higher TPR and lower FPR value (the optimal threshold ROC coordinates are shown in Figure 9). Furthermore, statistical analysis of the CRP classification results was performed based on the optimal segmentation threshold, as shown in Figure 10. The predicted results based on the four models showed that in the high-latitude and high-longitude regions, the majority of CRP pixel values were 0 (no coniferous sapling regeneration), while in the low-latitude and low-longitude regions, the majority of pixel values were 1 (coniferous sapling regeneration).

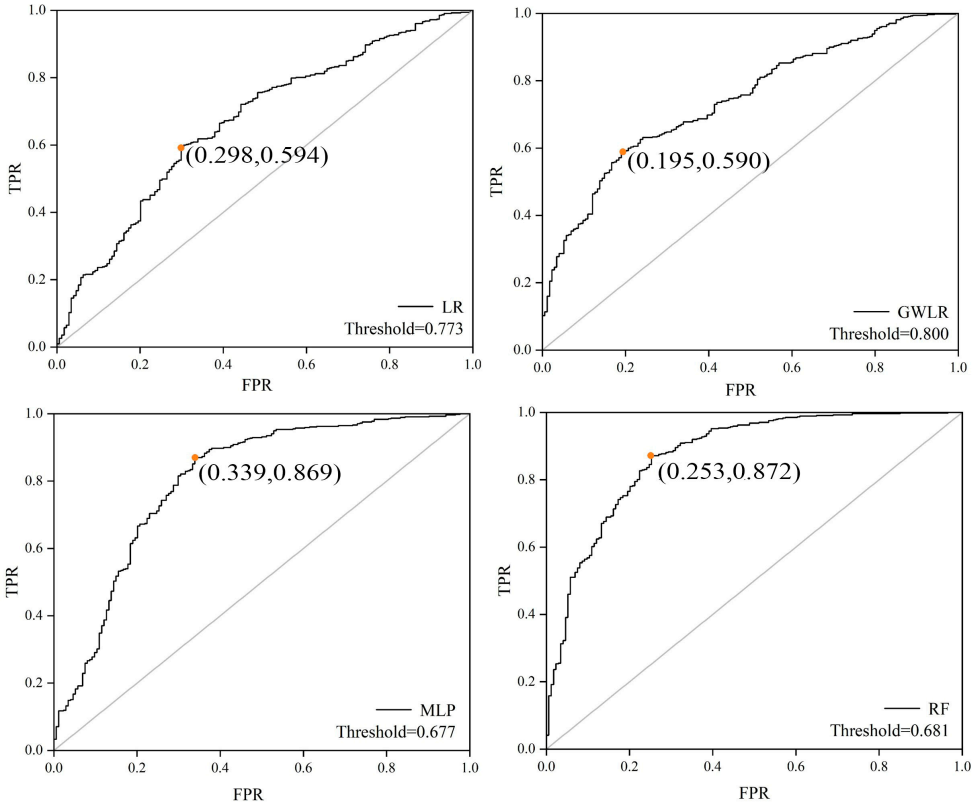

Figure 9. Determination of optimal threshold segmentation points for all samples. The points in the graphs are the ROC curve coordinates corresponding to the optimal segmentation thresholds.

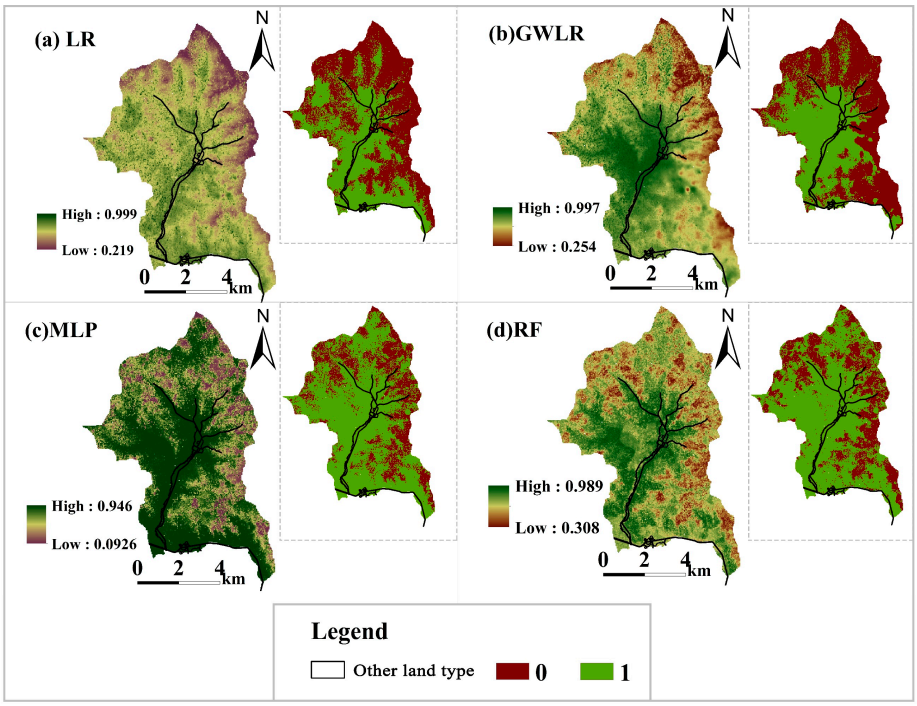

Figure 10. Spatial statistics of model prediction results. The large image shows the distribution trend of the CRP in the study area; the small image shows the segmentation results based on the optimal threshold; 0—no coniferous sapling regeneration; 1—coniferous sapling regeneration.

4. Discussion

To understand the size and distribution of the CRP in the Liangshui National Nature Reserve, we constructed LR, GWLR, MLP, and RF models for research and analysis. The four models had higher CRP prediction results in the low-latitude and low-longitude regions of the study area. Lower values of the CRP were obtained in the high-latitude and high-longitude regions of the study area. Coniferous sapling regeneration mainly occurred in the low-latitude and low-longitude regions of the study area, and in the high-latitude and high-longitude regions of the study area, most pixels had a CRP value of 0 (i.e., no coniferous sapling regeneration occurred). Based on the above research results, the following detailed analysis was conducted: (I) model variable selection; (II) selection of predictor variables and their ecological implications; (III) model comparison; and (IV) determination of the advantages of optimal threshold segmentation.

4.1. Model Variable Selection

In this study, a total of 43 variable factors were extracted. The data mainly consisted of measured data, remote sensing factors, and terrain factors. When studying understory regeneration, many scholars have explored and analyzed the relationship between stand factors and sapling regeneration. Hai jiao Yang et al. [51] found that moderate thinning, which controls stand density, can adjust the diversity of understory plants. Feng Liu et al. [9] used a nonlinear mixed model to analyze the impact of different sizes of forest gaps on the biomass accumulation of understory saplings. HH Chen et al. [10] analyzed the relationship between stand types and sapling regeneration in oak secondary forests. Maitane Erdozai et al. [52] studied the effects of forest thinning and climate on understory regeneration. However, we believe that relying solely on forest stand variable factors for forest regeneration research is not sufficient. The fitting accuracy of the models we constructed indicated that the LR and GWLR models constructed using the six variable factors screened by stepwise regression had a lower accuracy than the MLP and RF models constructed using all variables. Based on the LR and RF models, the DEM is the largest variable factor affecting CRP. Therefore, it is necessary to consider more remote sensing factor variables.

4.2. Selected Predictor Variables and Their Ecological Implications

We can see from the analysis of the LR and RF models that the DEM was the most important variable for predicting CRP. In this study, there was a negative correlation between the DEM and CRP based on the LR and GWLR models, indicating that the probability of coniferous sapling regeneration gradually decreased with increasing altitude. Related studies have shown a negative correlation between the DEM and species richness [53], and sapling species were also considered when calculating species richness. We speculate that this is the reason why the DEM variable is the most important and negatively correlated with the CRP in predicting the CRP.

Based on the fitting results of the LR model, the average DBH (AD, cm) variable was the second variable factor that affected the CRP. The fitting results of the LR model indicate a positive relationship between the average DBH (AD, cm) variables and the CRP. When the DBH of a single tree is large, the tree will have greater competitiveness [54], which can lead to the death of other trees due to insufficient nutritional space. When this situation occurs, forest density significantly decreases. As previously reported by scholars, the regeneration of saplings in secondary forests and pine oak forests is greatly affected by stand density [7,9]. We speculate that the regeneration of coniferous saplings in the natural forest we studied was related to stand density. Therefore, this may be the reason for ranking the average DBH (AD, cm) variables second in CRP prediction based on the LR model used in this study.

According to the importance ranking of the RF model, the volume of living trees per hectare (VLT, m^3/ha) variable ranked second. In addition, according to the GWLR model parameter coefficient interpolation results (Figure 7), the VLT variable factor had a negative correlation with the CRP in most of the study areas, and the absolute value of

the coefficient was greater in high-latitude and high-longitude areas than in most areas of the average DBH (AD, cm) variable factor. In high-latitude and high-longitude areas, the VLT variable may have a greater impact on the CRP than the average DBH (AD, cm) variable. Furthermore, the VLT is usually a variable used by scholars to study forest carbon storage [55,56], as there is a synergistic effect between forest carbon storage and forest regeneration [57]. Therefore, we speculate that this is the reason why the VLT ranked second in importance in the RF model.

4.3. Model Comparison

The accuracy of the RF and MLP models we built was greater than that of the LR and GWLR models. However, as we discussed above, the distribution of CRP at different latitudes and longitudes exhibited spatial heterogeneity. The MLP and RF models are nonparametric models, and when constructed, the RF and MLP models only consider the connections between data and do not consider the spatial heterogeneity between sample plots. This resulted in significant errors in the prediction of other sample points by the RF and MLP constructed with strong spatial heterogeneity between sample points, resulting in a significant decrease in model accuracy. The GWLR model constructed in this study can effectively address the issue of spatial heterogeneity between sample plots. To observe the spatial heterogeneity reduction effect of the GWLR model compared to the RF, MLP, and LR models, we used GeoDa 1.12.1 software to plot the Moran's I exponential coordination curve for model residuals (Figure 11). The lower the slope of the curve was, the smaller the residual Moran's I index value of the model. According to Figure 11, the values of the residual Moran's I index, the models could be sorted as the following: GWLR < MLP < RF < LR. The smaller the Moran's I index value of the model residual was, the more spatial factors were considered in the construction of the model. The GWLR had the lowest Moran's I index value compared to the other three models, meaning that the GWLR model considered more spatial factors during construction, further demonstrating that the LR, RF, and MLP models had weaker performance in reducing spatial autocorrelation between sample points than the GWLR model. However, the results of our model construction suggested that the accuracies of the RF and MLP models were greater than those of the GWLR model, which also reflected the advantages of the RF and MLP models. In this study, because the RF and MLP models sought connections between data without limiting the number of variables, we used 43 variable factors to construct the RF and MLP models to improve the accuracy of CRP prediction, while the GWLR model used only six variable factors. We speculate that this difference in the number of variables may be the main reason for the accuracy differences between the RF, MLP, and GWLR models. Finally, based on the results of our analysis, we infer that the RF and MLP models can be selected when there are many independent variable factors in the study of CRP. When there are few independent variable factors and high spatial heterogeneity between sample plots, the GWLR model can be selected for the prediction and analysis of the CRP.

4.4. Advantages of Optimal Threshold Segmentation

Few scholars have specified their classification thresholds when constructing classification models [18]. The LR model is an "s" curve, with 0.5 being the inflection point of the curve, so most scholars default to 0.5 as the classification threshold when constructing the LR model. We believe that doing so is very unreasonable. We plotted the kappa coefficients corresponding to different classification thresholds for the four models (Figure 12). Figure 12 shows that the segmentation threshold calculated using full sample data had different kappa values, indicating that the selection of the threshold had a significant impact on the classification accuracy. However, Figure 12 also shows that the optimal threshold selected in this study was not the maximum kappa point, and the corresponding kappa values were not significantly different from the maximum kappa values because the optimal thresholds were usually selected based on the ROC curves. The optimal threshold we chose was the point located at the top left corner of the ROC curve (i.e., the point with

the largest TPR-FPR value). In addition, most scholars use the classification algorithm of the RF model [58,59], but the classification algorithm predicts a CRP with only 0 or 1 results, without a specific probability value. In this study, we chose the RF regression model to predict the specific CRP size of each pixel, and based on the predicted results, we used the optimal classification threshold to perform 0–1 classification statistics. The advantage of this threshold was that it not only had a specific CRP value for each pixel, but also allowed for 0–1 classification statistics for each pixel in the research area. In summary, the advantage of choosing the optimal segmentation threshold is to reduce the false-positive rate (FPR) and improve the true-positive rate (TPR), thereby improving classification accuracy.

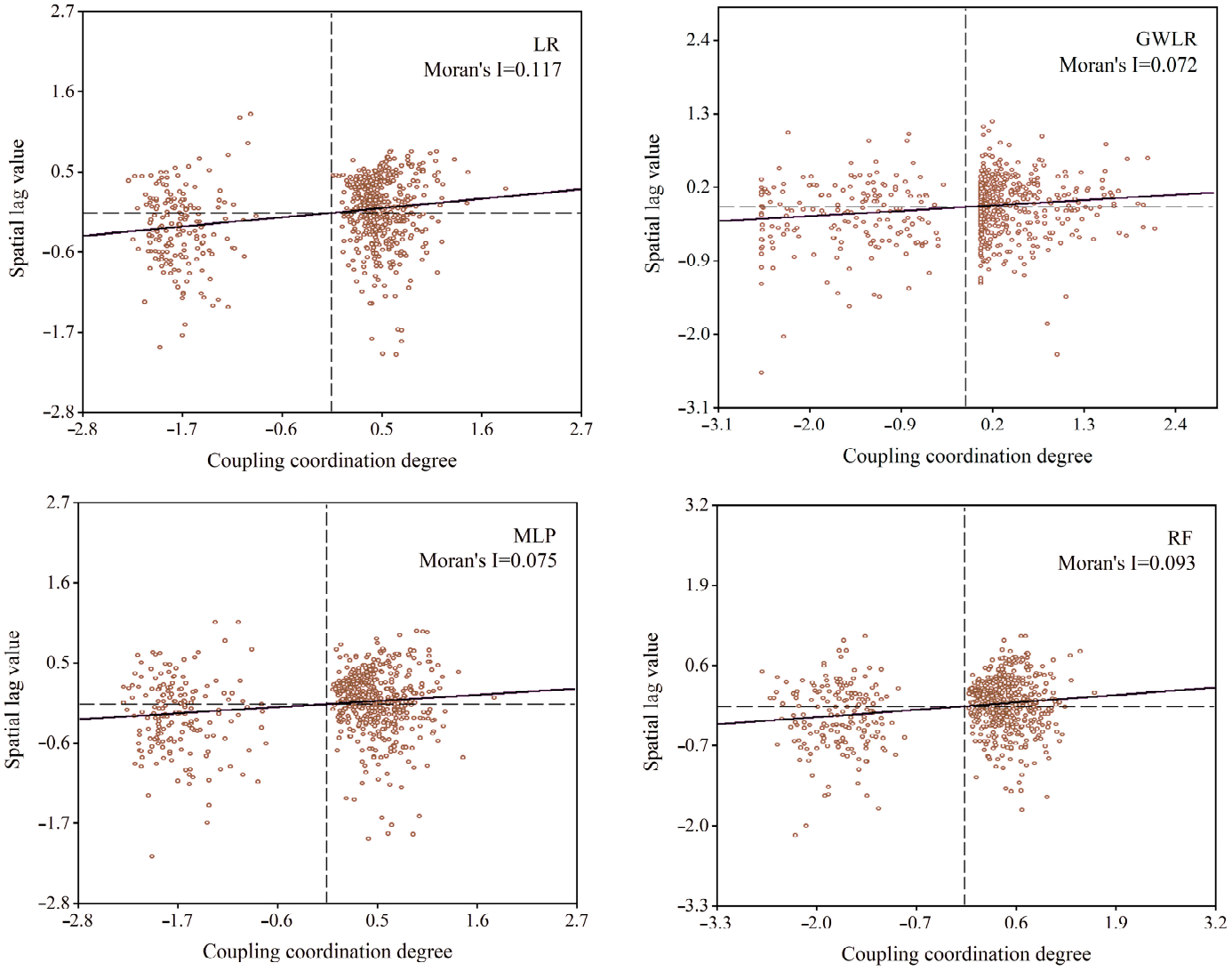

Figure 11. Moran's I coordination curves. The scatter plots in the figure represent the sample locations, and the curves were fitted based on them.

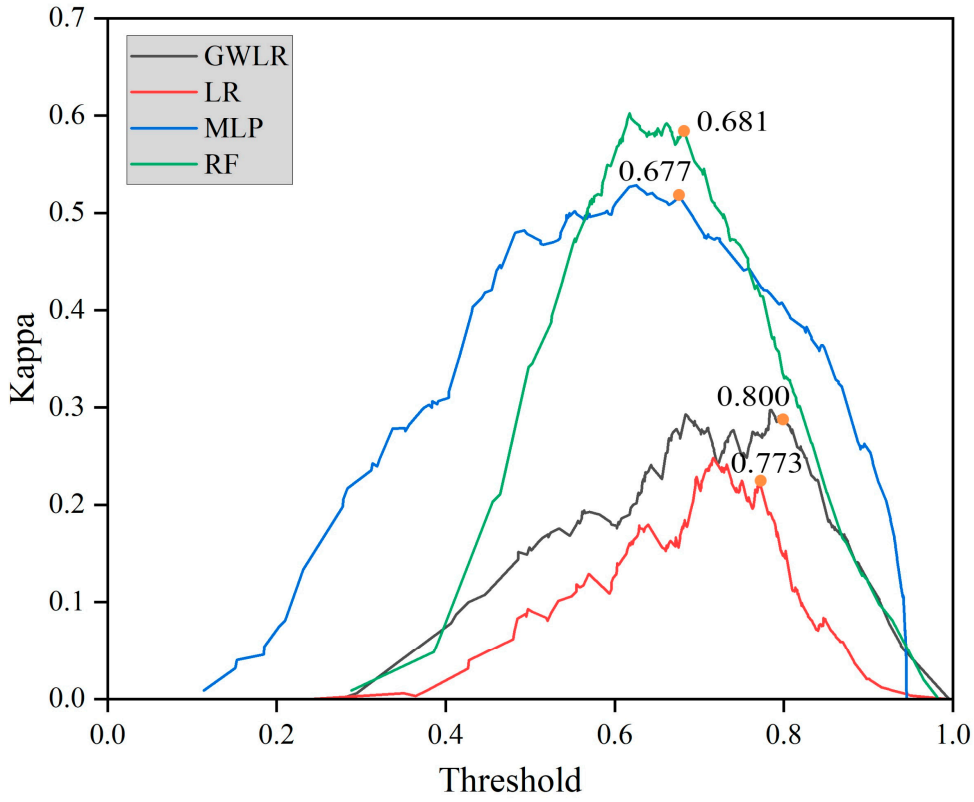

Figure 12. Kappa values corresponding to different segmentation thresholds.

5. Conclusions

Coniferous tree species are an important type of tree species in the Liangshui National Nature Reserve. Understanding the regeneration of coniferous saplings under their forests is key to predicting future forest structures and forest management. This study was based on forestry multisource remote sensing data, combined with field survey data, and LR, GWLR, RF, and MLP models were constructed. We drew the following conclusions:

1. The RF model achieved the highest value of accuracy evaluation. However, the RF model has the disadvantage of neglecting the spatial autocorrelation among neighboring samples. The GWLR model, constructed by LR regression, effectively accounts for the spatial autocorrelation among neighboring samples.
2. The distribution of CRP along the latitude and longitude lines exhibited spatial heterogeneity.
3. The DEM variable was the most significant factor influencing CRP.
4. Coniferous sapling regeneration mainly occurred in low-latitude and low-longitude regions, and most pixels in the high-latitude and high-longitude regions of the study had a CRP value of 0, indicating that no coniferous sapling regeneration occurred.

Author Contributions: Conceptualization, H.Z., Y.S. and W.J.; methodology, H.Z.; software, H.Z. and Z.Z.; validation, H.Z., Y.S. and F.W.; formal analysis, H.Z.; investigation, H.Z., Z.Z., S.W. and Y.S.; data curation, H.Z. and Y.S.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z. and Y.S.; supervision, W.J.; project administration, W.J.; funding acquisition, W.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China National Key Research and Development Program (Grant No. 2022YFD2201003-02) and the Special Fund Project for Basic Research in Central Universities (2572019CP08).

Data Availability Statement: Not applicable.

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

References

1. Haq, S.M.; Amjad, M.S.; Waheed, M.; Bussmann, R.W.; Proćków, J. The floristic quality assessment index as ecological health indicator for forest vegetation: A case study from Zabarwan Mountain Range, Himalayas. *Ecol. Indic.* **2022**, *145*, 109670. [CrossRef]
2. Constant, N.L.; Taylor, P.J. Restoring the forest revives our culture: Ecosystem services and values for ecological restoration across the rural-urban nexus in South Africa. *For. Policy Econ.* **2020**, *118*, 102222. [CrossRef]
3. de Pater, C.; Verschuuren, B.; Elands, B.; van Hal, I.; Turnhout, E. Spiritual values in forest management plans in British Columbia and the Netherlands. *For. Policy Econ.* **2023**, *151*, 102955. [CrossRef]
4. Taye, F.A.; Folkersen, M.V.; Fleming, C.M.; Buckwell, A.; Mackey, B.; Diwakar, K.C.; Le, D.; Hasan, S.; Ange, C.S. The economic values of global forest ecosystem services: A meta-analysis. *Ecol. Econ.* **2021**, *189*, 107145. [CrossRef]
5. Hammond, M.E.; Pokorný, R.; Okae-Anti, D.; Gyedu, A.; Obeng, I.O. The composition and diversity of natural regeneration of tree species in gaps under different intensities of forest disturbance. *J. For. Res.* **2021**, *32*, 1843–1853. [CrossRef]
6. Aide, T.M.; Zimmerman, J.K.; Pascarella, J.B.; Rivera, L.; Ecology, H.M.V.J.R. Forest Regeneration in a Chronosequence of Tropical Abandoned Pastures: Implications for Restoration Ecology. *Restor. Ecol.* **2000**, *8*, 328–338. [CrossRef]
7. Maciel-Nájera, J.F.; Hernández-Velasco, J.; González-Elizondo, M.S.; Hernández-Díaz, J.C.; López-Sánchez, C.A.; Antúnez, P.; Bailón-Soto, C.E.; Wehenkel, C. Unexpected spatial patterns of natural regeneration in typical uneven-aged mixed pine-oak forests in the Sierra Madre Occidental, Mexico. *Glob. Ecol. Conserv.* **2020**, *23*, e01074. [CrossRef]
8. Boag, A.E.; Ducey, M.J.; Palace, M.W.; Hartter, J. Topography and fire legacies drive variable post-fire juvenile conifer regeneration in eastern Oregon, USA. *For. Ecol. Manag.* **2020**, *474*, 118312. [CrossRef]
9. Liu, F.; Tan, C.; Yang, Z.; Li, J.; Xiao, H.; Tong, Y. Regeneration and growth of tree seedlings and saplings in created gaps of different sizes in a subtropical secondary forest in southern China. *For. Ecol. Manag.* **2022**, *511*, 120143. [CrossRef]
10. Chen, H.H.; Zhu, X.; Zhu, G.Y.; Liu, F.H. Effects of stand structure on understory biomass of the Quercus spp secondary forests in Hunan Province, China. *J. Appl. Ecol.* **2020**, *31*, 349–356.
11. Hu, T.; Sun, Y.; Jia, W.; Li, D.; Zou, M.; Zhang, M. Study on the Estimation of Forest Volume Based on Multi-Source Data. *Sensors* **2021**, *21*, 7796. [CrossRef] [PubMed]
12. Lee, L.X.; Whitby, T.G.; Munger, J.W.; Stonebrook, S.J.; Friedl, M.A. Remote sensing of seasonal variation of LAI and fAPAR in a deciduous broadleaf forest. *Agric. For. Meteorol.* **2023**, *333*, 109389. [CrossRef]
13. Iglseder, A.; Immitzer, M.; Dostálová, A.; Kasper, A.; Pfeifer, N.; Bauerhansl, C.; Schöttl, S.; Hollaus, M. The potential of combining satellite and airborne remote sensing data for habitat classification and monitoring in forest landscapes. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103131. [CrossRef]
14. Puliti, S.; Breidenbach, J.; Schumacher, J.; Hauglin, M.; Klingenberg, T.F.; Astrup, R. Above-ground biomass change estimation using national forest inventory data with Sentinel-2 and Landsat. *Remote Sens. Environ.* **2021**, *265*, 112644. [CrossRef]
15. Persson, H.J.; Ekström, M.; Ståhl, G. Quantify and account for field reference errors in forest remote sensing studies. *Remote Sens. Environ.* **2022**, *283*, 113302. [CrossRef]
16. Tariq, S.; Nawaz, H.; Mehmood, U.; ul Haq, Z.; Pata, U.K.; Murshed, M. Remote sensing of air pollution due to forest fires and dust storm over Balochistan (Pakistan). *Atmos. Pollut. Res.* **2023**, *14*, 101674. [CrossRef]
17. Stahl, A.T.; Andrus, R.; Hicke, J.A.; Hudak, A.T.; Bright, B.C.; Meddens, A.J.H. Automated attribution of forest disturbance types from remote sensing data: A synthesis. *Remote Sens. Environ.* **2023**, *285*, 113416. [CrossRef]
18. Kumar, R.; Nandy, S.; Agarwal, R.; Kushwaha, S.P.S. Forest cover dynamics analysis and prediction modeling using logistic regression model. *Ecol. Indic.* **2014**, *45*, 444–455. [CrossRef]
19. Basu, T.; Das, A.; Pereira, P. Exploring the drivers of urban expansion in a medium-class urban agglomeration in India using the remote sensing techniques and geographically weighted models. *Geogr. Sustain.* **2023**, *4*, 150–160. [CrossRef]
20. Fotheringham, A.S.; Charlton, M.E.; Brunsdon, C. Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis. *Environ. Plan. A Econ. Space* **2016**, *30*, 1905–1927. [CrossRef]
21. Sun, Y.; Ao, Z.; Jia, W.; Chen, Y.; Xu, K. A geographically weighted deep neural network model for research on the spatial distribution of the down dead wood volume in Liangshui National Nature Reserve (China). *Iforest-Biogeosci. For.* **2021**, *14*, 353–361. [CrossRef]
22. Brunsdon, C.; Fotheringham, A.S.; Charlton, M.E. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geogr. Anal.* **2010**, *28*, 281–298. [CrossRef]
23. Cui, Y.; Pan, C.; Liu, C.; Luo, M.; Guo, Y. Spatiotemporal variation and tendency analysis on rainfall erosivity in the Loess Plateau of China. *Hydrol. Res.* **2020**, *51*, 1048–1062. [CrossRef]
24. Liu, D.; Clarke, K.C.; Chen, N. Integrating spatial nonstationarity into SLEUTH for urban growth modeling: A case study in the Wuhan metropolitan area. *Comput. Environ. Urban Syst.* **2020**, *84*, 101545. [CrossRef]
25. Ali, M.R.; Nipu, S.M.A.; Khan, S.A. A decision support system for classifying supplier selection criteria using machine learning and random forest approach. *Decis. Anal. J.* **2023**, *7*, 100238. [CrossRef]
26. Ghosh, A.; Dey, P. Flood Severity assessment of the coastal tract situated between Muriganga and Saptamukhi estuaries of Sundarban delta of India using Frequency Ratio (FR), Fuzzy Logic (FL), Logistic Regression (LR) and Random Forest (RF) models. *Reg. Stud. Mar. Sci.* **2021**, *42*, 101624. [CrossRef]
27. Billah, M.; Islam, A.K.M.S.; Mamoon, W.B.; Rahman, M.R. Random forest classifications for landuse mapping to assess rapid flood damage using Sentinel-1 and Sentinel-2 data. *Remote Sens. Appl. Soc. Environ.* **2023**, *30*, 100947. [CrossRef]

28. Zermane, A.; Mohd Tohir, M.Z.; Zermane, H.; Baharudin, M.R.; Mohamed Yusoff, H. Predicting fatal fall from heights accidents using random forest classification machine learning model. *Saf. Sci.* **2023**, *159*, 106023. [CrossRef]
29. Karimi, B.; Hashemi, S.H.; Aghighi, H. Development of the best retrieval models of non-optically active parameters for an artificial shallow lake by random forest algorithm. *Remote Sens. Appl. Soc. Environ.* **2023**, *29*, 100926. [CrossRef]
30. Ghazvini, M.; Varedi-Koulaei, S.M.; Ahmadi, M.H.; Kim, M. Optimization of MLP neural network for modeling flow boiling performance of Al2O3/water nanofluids in a horizontal tube. *Eng. Anal. Bound. Elem.* **2022**, *145*, 363–395. [CrossRef]
31. Martínez-Comesaña, M.; Ogando-Martínez, A.; Troncoso-Pastoriza, F.; López-Gómez, J.; Febrero-Garrido, L.; Granada-Álvarez, E. Use of optimised MLP neural networks for spatiotemporal estimation of indoor environmental conditions of existing buildings. *Build. Environ.* **2021**, *205*, 108243. [CrossRef]
32. Wang, F.; Sun, Y.; Jia, W.; Zhu, W.; Li, D.; Zhang, X.; Tang, Y.; Guo, H. Development of Estimation Models for Individual Tree Aboveground Biomass Based on TLS-Derived Parameters. *Forests* **2023**, *14*, 351. [CrossRef]
33. Li, H.; Zhang, G.; Zhong, Q.; Xing, L.; Du, H. Prediction of Urban Forest Aboveground Carbon Using Machine Learning Based on Landsat 8 and Sentinel-2: A Case Study of Shanghai, China. *Remote Sens.* **2023**, *15*, 284. [CrossRef]
34. Meng, X.; Li, F. *Forest Mensuration*, 3rd ed.; China Forestry Publishing House: Beijing, China, 2006; pp. 28, 59, 61, 63, 67.
35. Vanderhoof, M.K.; Alexander, L.; Christensen, J.; Solvik, K.; Nieuwlandt, P.; Sagehorn, M. High-frequency time series comparison of Sentinel-1 and Sentinel-2 satellites for mapping open and vegetated water across the United States (2017–2021). *Remote Sens. Environ.* **2023**, *288*, 113498. [CrossRef]
36. Liu, X.; Frey, J.; Munteanu, C.; Still, N.; Koch, B. Mapping tree species diversity in temperate montane forests using Sentinel-1 and Sentinel-2 imagery and topography data. *Remote Sens. Environ.* **2023**, *292*, 113576. [CrossRef]
37. Sandhini Putri, A.F.; Widyatmanti, W.; Umarhadi, D.A. Sentinel-1 and Sentinel-2 data fusion to distinguish building damage level of the 2018 Lombok Earthquake. *Remote Sens. Appl. Soc. Environ.* **2022**, *26*, 100724. [CrossRef]
38. Collins, M.J.; Wiebe, J.; Clausi, D.A. The effect of speckle filtering on scale-dependent texture estimation of a forested scene. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 1160–1170. [CrossRef]
39. Prasad, T.S.; Gupta, R.K. Texture based classification of multidate SAR images—A case study. *Geocarto Int.* **1998**, *13*, 53–62. [CrossRef]
40. Desloires, J.; Ienco, D.; Botrel, A. Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. *Comput. Electron. Agric.* **2023**, *209*, 107807. [CrossRef]
41. Eskandari, S.; Ali Mahmoudi Sarab, S. Mapping land cover and forest density in Zagros forests of Khuzestan province in Iran: A study based on Sentinel-2, Google Earth and field data. *Ecol. Inform.* **2022**, *70*, 101727. [CrossRef]
42. Yang, X.; Qiu, S.; Zhu, Z.; Rittenhouse, C.; Riordan, D.; Cullerton, M. Mapping understory plant communities in deciduous forests from Sentinel-2 time series. *Remote Sens. Environ.* **2023**, *293*, 113601. [CrossRef]
43. Baret, F.; Guyot, G. Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote Sens. Environ.* **1991**, *35*, 161–173. [CrossRef]
44. Crippen, R.E. Calculating the vegetation index faster. *Remote Sens. Environ.* **1990**, *34*, 71–73. [CrossRef]
45. Li, W.; Li, C.; Jiang, L. Learning from crowds with robust logistic regression. *Inf. Sci.* **2023**, *639*, 119010. [CrossRef]
46. Karabadji, N.E.I.; Amara Korba, A.; Assi, A.; Seridi, H.; Aridhi, S.; Dhifli, W. Accuracy and diversity-aware multi-objective approach for random forest construction. *Expert Syst. Appl.* **2023**, *225*, 120138. [CrossRef]
47. Dong, Y.-H.; Peng, F.-L.; Li, H.; Men, Y.-Q. Spatial autocorrelation and spatial heterogeneity of underground parking space development in Chinese megacities based on multisource open data. *Appl. Geogr.* **2023**, *153*, 102897. [CrossRef]
48. Hoyos, N.; Escobar, J.; Restrepo, J.C.; Arango, A.M.; Ortiz, J.C. Impact of the 2010–2011 La Niña phenomenon in Colombia, South America: The human toll of an extreme weather event. *Appl. Geogr.* **2013**, *39*, 16–25. [CrossRef]
49. Young, S.G.; Jensen, R.R. Statistical and visual analysis of human West Nile virus infection in the United States, 1999–2008. *Appl. Geogr.* **2012**, *34*, 425–431. [CrossRef]
50. Moran, P.A.P. The Interpretation of Statistical Maps. *J. R. Stat. Soc. Ser. B Methodol.* **1948**, *10*, 243–251. [CrossRef]
51. Yang, H.; Pan, C.; Wu, Y.; Qing, S.; Wang, Z.; Wang, D. Response of understory plant species richness and tree regeneration to thinning in Pinus tabuliformis plantations in northern China. *For. Ecosyst.* **2023**, *10*, 100105. [CrossRef]
52. Erdozain, M.; Bonet, J.A.; Martínez de Aragón, J.; de-Miguel, S. Forest thinning and climate interactions driving early-stage regeneration dynamics of maritime pine in Mediterranean areas. *For. Ecol. Manag.* **2023**, *539*, 121036. [CrossRef]
53. Bruun, H.H.; Moen, J.; Virtanen, R.; Grytnes, J.A.; Oksanen, L.; Angerbjörn, A. Effects of altitude and topography on species richness of vascular plants, bryophytes and lichens in alpine communities. *J. Veg. Sci.* **2010**, *17*, 37–46. [CrossRef]
54. Qin, Y.; He, X.; Lei, X.; Feng, L.; Zhou, Z.; Lu, J. Tree size inequality and competition effects on nonlinear mixed effects crown width model for natural spruce-fir-broadleaf mixed forest in northeast China. *For. Ecol. Manag.* **2022**, *518*, 120291. [CrossRef]
55. Birungi, V.; Dejene, S.W.; Mbogga, M.S.; Dumas-Johansen, M. Carbon stock of Agoro Agu Central Forest reserve, in Lamwo district, Northern Uganda. *Heliyon* **2023**, *9*, e14252. [CrossRef] [PubMed]
56. Basyuni, M.; Wirasatriya, A.; Iryanthony, S.B.; Amelia, R.; Slamet, B.; Sulistiyono, N.; Pribadi, R.; Sumarga, E.; Eddy, S.; Al Mustaniroh, S.S.; et al. Aboveground biomass and carbon stock estimation using UAV photogrammetry in Indonesian mangroves and other competing land uses. *Ecol. Inform.* **2023**, *77*, 102227. [CrossRef]
57. Salete Capellesso, E.; Cequinel, A.; Marques, R.; Luisa Sausen, T.; Bayer, C.; Marques, M.C.M. Co-benefits in biodiversity conservation and carbon stock during forest regeneration in a preserved tropical landscape. *For. Ecol. Manag.* **2021**, *492*, 119222. [CrossRef]

58. Hart, E.; Sim, K.; Kamimura, K.; Meredieu, C.; Guyon, D.; Gardiner, B. Use of machine learning techniques to model wind damage to forests. *Agric. For. Meteorol.* **2019**, *265*, 16–29. [CrossRef]
59. Dobrini, D.; Gaparovi, M.; Medak, D.J.R.S. Sentinel-1 and 2 Time-Series for Vegetation Mapping Using Random Forest Classification: A Case Study of Northern Croatia. *Remote Sens.* **2021**, *13*, 2321. [CrossRef]