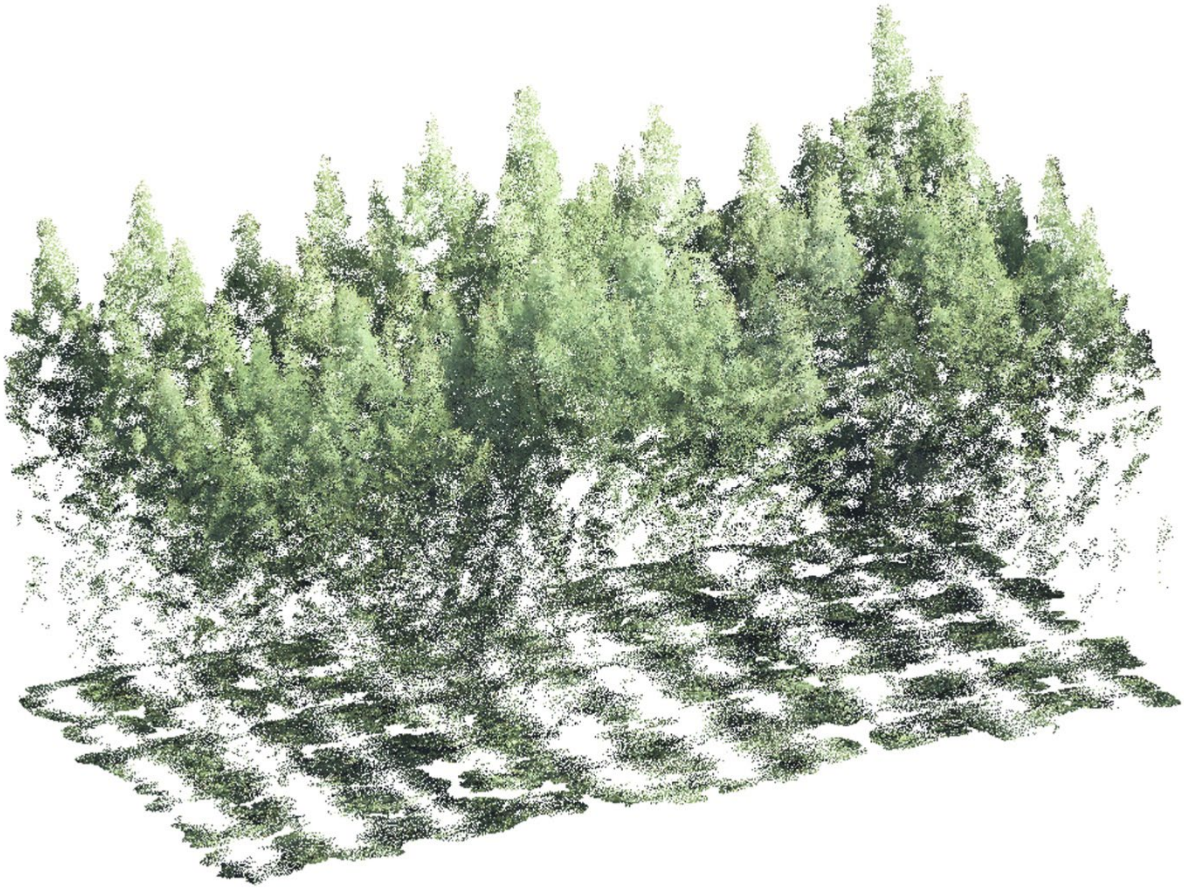


Estimating individual tree variables using UAV LiDAR

Zhaoying Pan

Supervised by: Dr. Vega Xu and Dr. Trevor Best

NZ School of Forestry, Christchurch, New Zealand

University of Canterbury

Acknowledgements

I would like to thank my primary supervisor Dr. Cong Xu for his valuable guidance and advice throughout the research project. Secondly, I would like to thank my associate supervisor Dr. Ning Ye for his technical support and academic advice. Many thanks to Dr. Trevor Best, Prof. Rien Visser and Dr. Campbell Harvey for their valuable advice.

Abstract

The use of Light Detection and Ranging (LiDAR) in forest inventory is increasing and is expected to continue rising. Combining unmanned aerial vehicles (UAV) with LiDAR can efficiently generate high-resolution point cloud data. By analysing these point cloud data, metrics at different levels can be extracted to analyse variables of individual trees within forests.

This report uses point cloud data generated by LiDAR, analyses it through a combination of algorithms, generates a normalized Canopy Height Model (nCHM), and completes individual tree detection and extraction. The report employs both automatic individual tree segmentation results and manually corrected results to generate models, comparing the prediction accuracy of models produced by these two segmentation methods. Linear models and Random Forest (RF) algorithms are generated in this report to predict individual tree height, diameter at breast height (DBH), and volume, with a total of 12 models being produced.

Through comparison, traditional linear models exceeded machine learning models in prediction accuracy for all three variables. Among these, the prediction results for tree height, DBH, and volume explained 82.06%, 58.89%, and 71.91% of the dependent variable variance, with absolute means of 2.97%, 5.55%, and 11.72% respectively. The results indicate that the prediction models for individual tree height and volume have a high degree of fit. Furthermore, the prediction results for individual tree height and volume suggest that models based on UAV LiDAR are capable of replacing formula-based individual tree volume estimation. However, the relatively low accuracy of DBH prediction indicates that challenges remain in estimating this parameter. Future research should focus on improving DBH prediction accuracy and optimising automatic segmentation algorithms.

Table of Contents

Acknowledgements 2
Abstract 3
Introduction 5
Methodology 7
Study area description 7
Software Used 7
LiDAR Data Collection 8
Collection of Dependent Variables 9
LiDAR Data Processing 9
Result 15
Tree Height Prediction Model 15
DBH Prediction Model 16
Volume Prediction Model 18
Discussion 19
Tree height Prediction Model Results Discussion 19
DBH Prediction Model Results Discussion 19
DBH Prediction Model Diagnostic Results 20
Volume Prediction Model Results Discussion 22
Volume Prediction Model Diagnostic Results 23
Future Study 25
Conclusion 25
References 27
Appendix 29

Introduction

Light Detection and ranging (LiDAR) technology, as a form of remote sensing (RS), is becoming increasingly important in forest management and monitoring in New Zealand. LiDAR can be mounted on various platforms to meet different needs. Installing LiDAR on unmanned aerial vehicles (UAVs) for scanning and data acquisition offers more flexibility and lower costs, making it considered a suitable LiDAR platform for forest environments. Over the past decade, the utilization rate of LiDAR has significantly increased from 17% in 2013 to 93% in 2023, with usage expected to continue rising (Manning, 2023).

UAV LiDAR evolved from Airborne Laser Scanning (ALS) and was first applied to tree measurements in 2010 (Anttoni et al, 2010). UAVs are divided into two types: rotary-wing and fixed-wing. Their applications in forestry include resource inventory, disease mapping, species classification, fire monitoring and impact assessment, quantifying spatial gaps, and estimating soil displacement after logging (Torresan et al, 2017). Compared to traditional RS platforms, UAV LiDAR platforms offer advantages in flexibility, speed, and cost-effectiveness (Dunford, 2009). Additionally, UAV LiDAR can collect reliable and dense 3D point data, enabling higher precision forest measurements (Cao et al., 2019).

New Zealand forestry commonly uses LiDAR technology for Area-Based Analysis (ABA) in forest inventory. ABA divides forests into fixed-size grids, analysing forest characteristics within each grid to calculate parameters such as average tree height, mean diameter at breast height, and average volume. ABA can accurately predict average height, basal area, mean volume, and biomass (Yu et al., 2010). Mei et al. (2023) used point cloud percentage resampling tools to thin original point cloud density, proving that ABA's dependence on point density is relatively low, and higher point density cannot further improve ABA's variable prediction accuracy.

ABA results primarily aim to provide averages or totals for all trees within a grid. As ABA divides forests into fixed-size grids, it cannot accurately reflect tree aggregation or sparse areas when these grids have spatial heterogeneity. Additionally, ABA cannot precisely analyse tree size, shape, and health conditions. These limitations lead to reduced accuracy in obtaining ecological and economic information, resulting in suboptimal management decisions.

With the application of UAV LiDAR in forestry, ABA's inability to improve analysis accuracy through high-resolution point cloud data has gradually become a bottleneck in forest inventory. How to conduct more precise forest inventory using high-resolution point cloud data has become a new development trend, creating potential opportunities for Tree-Based Analysis (TBA).

As UAVs can fly at a constant speed and height, they can obtain relatively high point cloud density. Moreover, the cost of obtaining high-density point cloud information via UAV is low, only requiring multiple charges and flights of the UAV. Torresan et al. (2020) compared the results of the itcLiDAR algorithm and li2012 algorithm in R, showing that the CHM-based itcLiDAR algorithm is more accurate for forests composed of conifers with distinct tops, consistent heights, and single layers compared to the point cloud-based li2012 algorithm. However, when facing two-layered dense mixed forests, due to point density limitations (193 points/m²), individual tree crown segmentation using only UAV LiDAR data cannot achieve satisfactory results. Michael et al. (2024) studied the point cloud density required for accurate TBA, finding that RMSE for DBH and volume stabilizes when point cloud density exceeds 400 to 750 points per square meter. The high pulse density laser data required for TBA has been greatly improved with the development of LiDAR sensors and UAV technology (Yu et al, 2010). Chisholm et al. (2021) used under-canopy UAVs to estimate tree DBH, finding high correlation between UAV-estimated and manually measured DBH in forests with larger tree sizes.

The exploration of Tree-Based Analysis (TBA) continues. While Area-Based Analysis (ABA) provides overall stand statistics, TBA can extract parameters for individual trees, allowing for more precise estimation of tree height, diameter, and volume (Hyyppä et al., 2012). The uncertainty in TBA estimation models is less dependent on plot size, allowing for calibration using individual trees and small plots (Dalponte). Shugart et al. (2015) suggest that TBA can track the growth conditions of individual trees, achieving continuous and high-frequency forest inventory. However, despite these advantages, TBA is inherently computation-intensive, and overcoming these complex calculations and achieving accurate tree segmentation are the main challenges in TBA development (Shugart et al., 2015).

In recent years, with the development of various algorithms, the rich information in LiDAR data has been further exploited. These algorithms can effectively classify and filter vast amounts of data, greatly reducing analysis time and improving efficiency. They can extract various metrics from LiDAR point cloud data, such as crown height, canopy density, and vertical structure characteristics, and then associate these indicators with field-measured tree parameters. By establishing this association, predictive models can be developed to quickly and accurately estimate individual tree parameters over large areas. Michael et al. (2024) categorized these metrics usable for model construction and precise calculations into three types: point-based, region-based, and pixel-based, all of which can be obtained through R software packages. Among these, height metrics (describing the statistical distribution of z-values in the point cloud), canopy metrics (providing measures of canopy structure), intensity metrics (describing the intensity of each pulse return), and other metrics (such as differences between height percentiles) are considered to contribute significantly to model construction (Xu et al., 2019; Cao et al., 2019).

Xu et al. (2019) compared four different model construction methods, where MLR and SUR are parametric models, while k-NN and Random Forest are non-parametric models. Results show that parametric and non-parametric models have little difference in volume estimation, both achieving relatively high coefficients of determination. MLR using LiDAR-derived indicators is considered the best modeling method for predicting stand variables in small-scale plantations in New Zealand.

The objective of this study is to create two models combining UAV LiDAR-derived indicators with field measurements to assess their prediction accuracy for individual radiata pine trees in terms of height, DBH, and volume. These two models will be based on parametric and non-parametric approaches respectively. After model establishment, they will be applied to predict the height, DBH, and volume of individual pine trees and compared with ground measurements to observe their fit. Simultaneously, the fit of the two types of models will be compared to determine which is more suitable for tree-based estimation of height, DBH, and volume (TBA). Additionally, this study will observe which indicators play more important roles in model prediction.

Methodology

Study area description

The study area is located in Rolleston (43°37'S, 172°21'E) on the South Island of New Zealand, with a forest area of approximately 8 ha. The experimental stand is a 16-year-old even-aged pure forest of radiata pine (Pinus radiata D. Don), with relatively flat terrain. Afforestation began in 2008, with radiata pine as the primary species. The study area includes three types of stockings: high density (2500 stems/ha), medium density (1250 stems/ha), and low density (625 stems/ha). Based on these stockings, the forest is divided into 48 square plots. In this study, 300 trees from five low-density plots were selected as sample trees, with the geographical overview of these five plots shown in Figure 1. The reason for choosing low-density plots as the research subject is that they more closely resemble New Zealand's commercial forests, making the research results more practical and valuable for wider application.

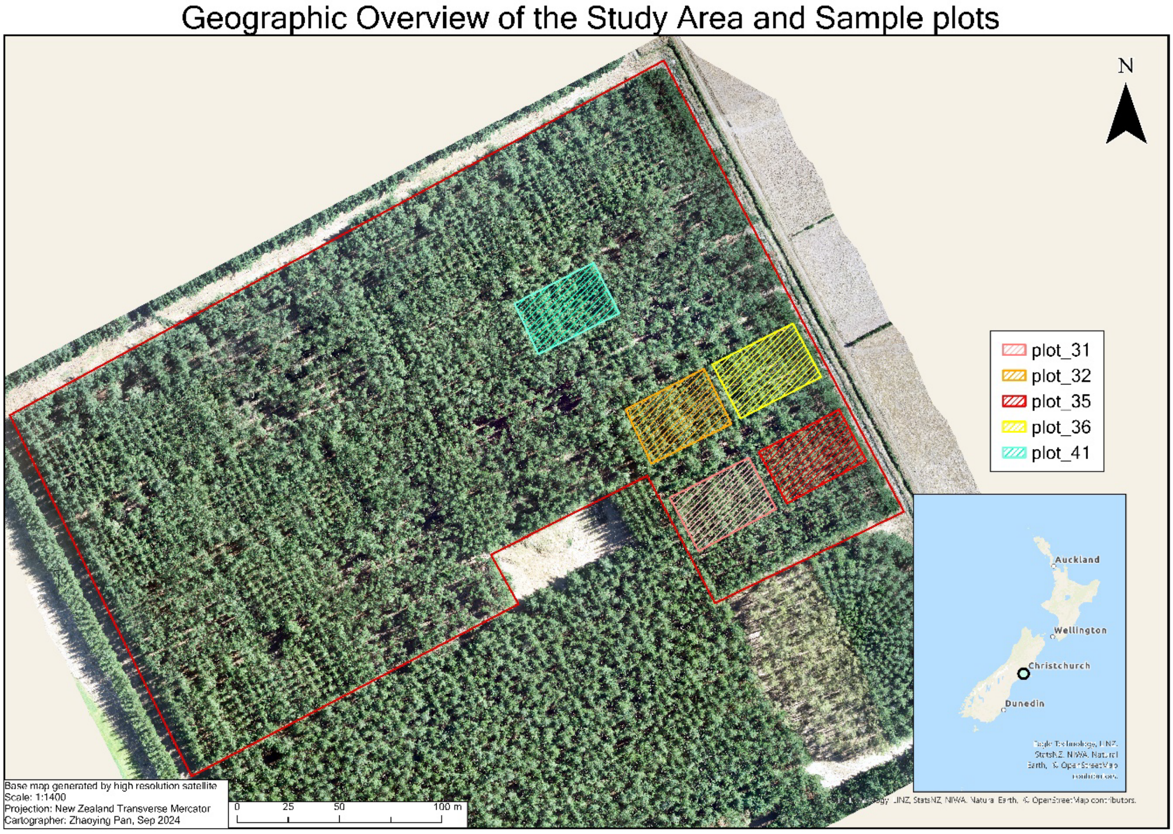

Geographic Overview of the Study Area and Sample plots
N
plot_31
plot_32
plot_35
plot_36
plot_41
Auckland
Wellington
Christchurch
Dunedin
Eagle Technology, LINZ, StatsNZ, NIWA, Natural Earth, © OpenStreetMap contributors.
Eagle Technology, LINZ, StatsNZ, NIWA, Natural Earth, © OpenStreetMap contributors.
Base map generated by high resolution satellite
Scale: 1:1400
Projection: New Zealand Transverse Mercator
Cartographer: Zhaoying Pan, Sep 2024
0
25
50
100 m

Figure 1. Geographic overview of study area and sample plots.

Software Used

The main software used in this project are CloudCompare, R, and ArcGIS Pro. CloudCompare is used for LiDAR data visualization, ArcGIS Pro is used for adjusting and checking constructed polygons and coordinate points of sample trees, and R will be combined with other packages for processing and analyzing LiDAR data, constructing models, training models, and model validation and evaluation. The data analysis and statistical modeling process in this study is completed using R language (R Core Team, 2024), with RStudio (Posit team, 2024) used as the integrated development environment to improve work efficiency.

LiDAR Data Collection

LiDAR data was collected in July 2023, with point cloud data visualization through CloudCompare shown in Figure 2. The equipment used for collecting LiDAR data was a Dji M300 RTK with L1 LiDAR solution, flying at a height of 80m above ground level, with 80% overlap, capturing triple returns at a sampling rate of 160Hz. This ensures that the LiDAR data has a relatively high resolution and achieves a balance between data quality and data processing workload. The LiDAR data collection covered a total area of 24 hectares, comprising approximately 200 million points, with a point density of 828 points/m² and a pulse density of 661.7 pulses/m². The preset coordinate system for the LiDAR data collection equipment was NZGD/New Zealand Transverse Mercator 2000.

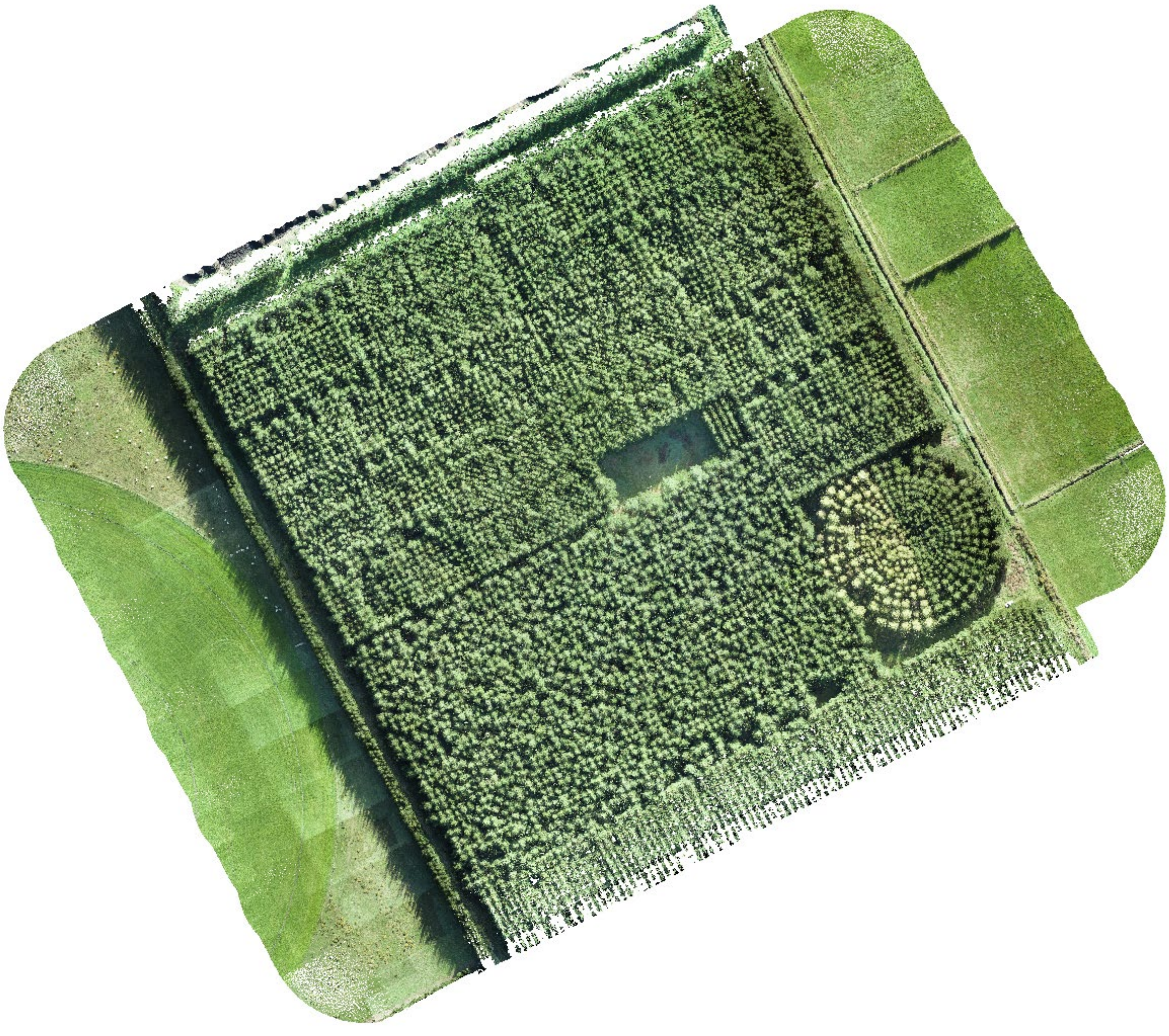


Figure 2. UAV LiDAR data visualisation results generated by cloudcompare

Collection of Dependent Variables

Sample tree information was measured in July 2023. In addition to measuring tree height using a Vertex hypsometer and DBH using a diameter tape, visual observations of the sample trees were also conducted to describe any abnormal conditions. Recording tree abnormalities played a crucial role in model construction and validation, as abnormal trees would lead to data that lacks referential value. If these trees were included as training samples in model construction, it would significantly decrease the prediction accuracy of the model itself. On the other hand, the model's prediction accuracy for trees would be underestimated. After excluding dead trees and trees with damaged crowns, a total of 284 trees were used for model training. The tree conditions in the five plots are shown in Table 1.

Table 1. Overview of the sample trees of each plot and the number of sample trees involved in model training

	plot31	plot32	plot35	plot36	plot41
Total number of trees	60	60	60	60	58
Broken Top	5	0	2	3	3
Dead Tree	0	0	0	0	1
Trees used in model training	55	60	58	57	54
sum				**284**	

The main reasons for measuring DBH and tree height, besides assessing tree growth conditions, include an important function: calculating tree volume. Currently, the widely used volume calculation formula is the method proposed by Kimberly & Beets (2007) for New Zealand radiata pine. The formula is as follows:

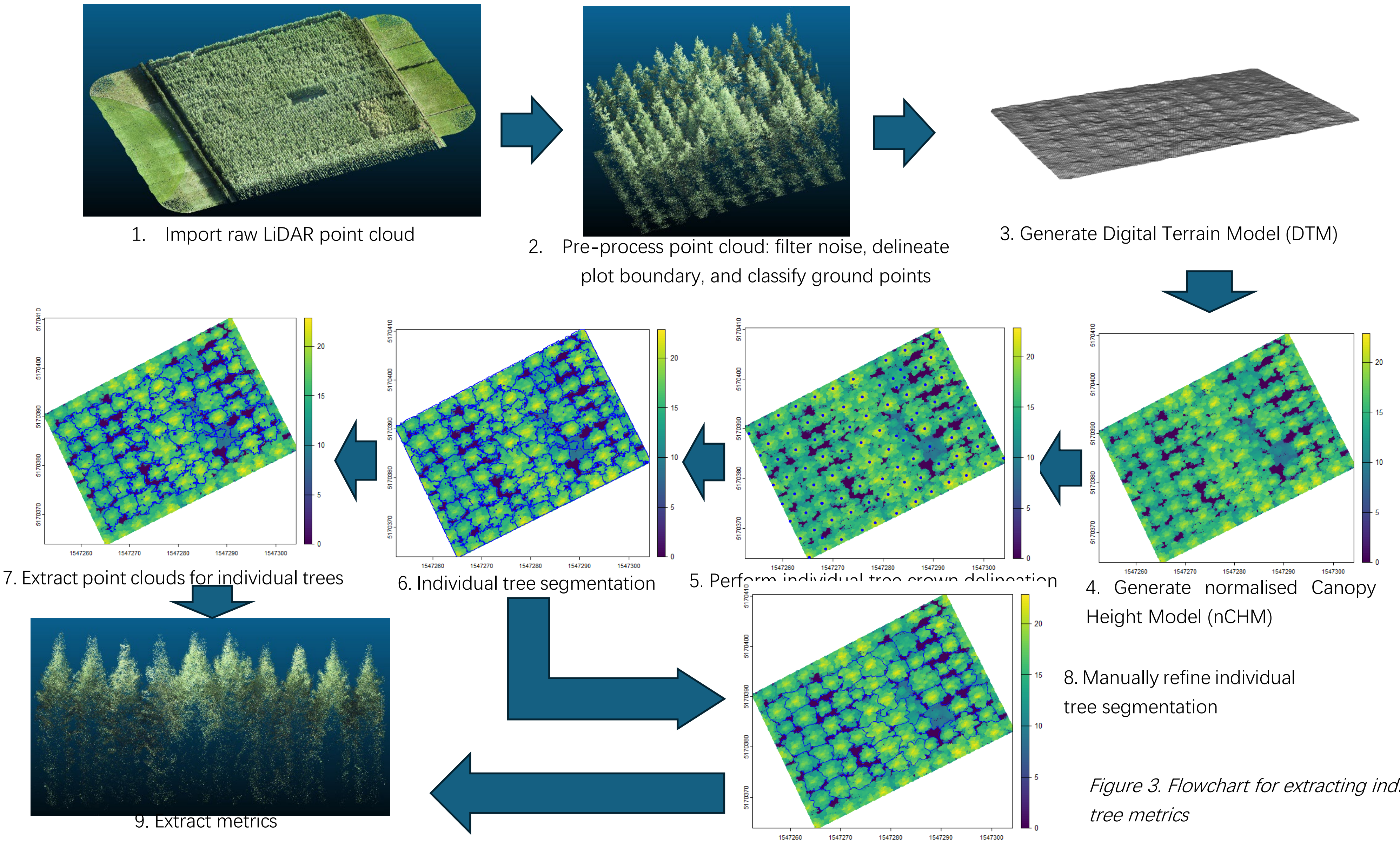
$$V = h \times \pi \times (\frac{DBH}{20})^2 \times [a \times (h - 1.4)^{-b} + c]$$

Where V is the tree volume (m^3), DBH is the diameter at breast height (mm), h is the tree height (m), a = 0.860, b = 0.972, c = 0.304.

In this project, the tree height and DBH measured in the field, as well as the tree volume calculated using the above formula, are used as standard values. These are compared with the predicted values from the model to determine the prediction accuracy of the model.

LiDAR Data Processing

The processing of LiDAR data includes nine steps, from importing the original point cloud data to extracting individual tree metrics. The workflow is shown in Figure 3. In the flowchart, the point cloud data processing for the five plots is repetitive, so plot 32 is used as an example for illustration.



1. Import raw LiDAR point cloud

2. Pre-process point cloud: filter noise, delineate plot boundary, and classify ground points

3. Generate Digital Terrain Model (DTM)

4. Generate normalised Canopy Height Model (nCHM)

5. Perform individual tree crown delineation

6. Individual tree segmentation

7. Extract point clouds for individual trees

8. Manually refine individual tree segmentation

9. Extract metrics

Figure 3. Flowchart for extracting individual tree metrics

Importing LAS Data and Extracting Target Plot Point Cloud Data

Processing and analysis of LiDAR point cloud data require installing and loading the lidR package in the R environment. After data import, a series of preprocessing steps were performed. First, the clip_roi function was used to extract the point cloud data of the target sample plot. Torresan et al. (2020) pointed out that when extracting the target sample plot, the edge effect needs to be considered. The edge effect can lead to two situations: tree trunks inside the plot with crowns outside, or trunks outside with parts of the crown entering the plot. The former can result in incomplete individual tree crown data, affecting model accuracy; the latter may lead to non-target trees being erroneously included. To mitigate the edge effect, a buffer zone of about 2.5 meters outside the plot boundary was set when extracting point cloud data to ensure complete capture of point cloud information for all target trees within the plot.

Preprocessing Point Cloud Data of the Target Area

After extracting the target plot point cloud, data cleaning was performed. This included removing duplicate points and outliers to improve data quality and analysis efficiency. UAV LiDAR typically uses a reciprocating scan mode, which may produce overlapping areas during data stitching, leading to point duplication. Additionally, outliers or noise points may occur during data collection, possibly due to sensor errors or interference from non-target objects (such as flying birds). These redundant or abnormal data points may affect the accuracy of subsequent analyses. By applying the filter_duplicates and las_filter functions from the lidR package, duplicate points and outliers were removed, generating a more reliable and high-quality point cloud dataset.

Ground Point Classification

The LiDAR point cloud dataset is essentially composed of a large number of unanalyzed and unclassified three-dimensional coordinate points. To conduct effective terrain analysis and vegetation feature extraction, ground points need to be identified. The lidR package provides multiple ground point classification algorithms, including the Progressive Morphological Filter (PMF) and Cloth Simulation Filter (CSF). The PMF algorithm is generally more suitable for areas with less terrain variation, while the CSF algorithm performs better in complex and varied terrains. Given that the study area has relatively flat terrain, the PMF algorithm was used for ground point classification. After completing ground point classification, the classify_noise function from the lidR package, combined with the Isolated Voxel Filter (IVF) method, was used to identify and classify noise points in the point cloud data. A voxel size of 5 meters was set, and areas with fewer than 6 points in each voxel were identified as noise. The IVF algorithm can remove abnormal points caused by equipment errors or environmental factors, thereby improving the accuracy of subsequent analyses.

Generating and Smoothing DTM

After completing ground point classification, each point in the point cloud data was assigned a "Classification" attribute. According to the LAS format standard, ground points were given a classification value of 2, while non-ground points were assigned a value of 1. Based on this classification, a Digital Terrain Model (DTM) can be generated. The generation of DTM mainly relies on the Triangulated Irregular Network (TIN) algorithm, which connects irregularly distributed ground points into a triangular network, forming a terrain surface composed of multiple triangular planes.

After generating the original DTM, it needs to be smoothed using the focal function from the raster package to reduce potential noise. This process helps eliminate local minor fluctuations, making the DTM better reflect overall terrain features while reducing subtle errors that may be introduced by data collection or processing.

Normalization

After generating the smoothed DTM, the next crucial step is to normalize the elevation of the point cloud data. This process is implemented using the normalize_height function from the lidR package. The main purpose of elevation normalization is to convert the absolute height of all points to height relative to the ground, which is vital for subsequent vegetation analysis. This step eliminates the impact of terrain undulations on vegetation height measurements and allows direct comparison of vegetation heights under different terrain conditions.

Generating CHM

Following the elevation normalization of point cloud data, the next key step is to generate a rasterized Canopy Height Model (CHM). CHM provides a visual representation of the forest's three-dimensional structure, where each pixel value represents the maximum vegetation height at that location. After filtering out points with heights less than 0 to eliminate underground points and noise data, the rasterize_canopy function is used to convert point cloud data into raster format. During the rasterization process, null values may occur due to the discrete nature of LiDAR data, so the K-Nearest Neighbor Inverse Distance Weighting (knnidw) algorithm is used for interpolation filling. This method considers the values and distances of surrounding known points, providing smooth results that take local features into account. To further improve CHM quality, a 3x3 pixel moving window average filter is applied to reduce noise and produce a continuous and smooth surface.

Individual Tree Detection

After generating the CHM, the next step is to use the ForestTools package for individual tree detection. ForestTools uses the Variable Window Filter (VWF) algorithm for tree top detection. In this study, a linear function is used to dynamically adjust the size of the detection window. This approach is suitable for handling radiata pine stands that may have double/multi leader issues. By adjusting the window size, the algorithm can more accurately identify the tops of trees at different heights, reducing the likelihood of mistaking multiple tops of the same tree for different trees. Based on the tree heights measured in the ground measurement for each plot, minimum heights for individual tree extraction were set separately, meaning the algorithm would ignore potential treetops below this height, helping to filter out shrubs or young trees. The optimal results of individual tree detection for the five plots and the raster images are presented in Appendix 1, with a detection accuracy of 99.65%.

Individual Tree Segmentation

Crown segmentation uses the marker-controlled watershed segmentation (MCWS) algorithm, utilizing detected treetops as markers and combining CHM height information to delineate crown boundaries. This method can effectively handle complex crown structures and closely connected trees. After completing crown segmentation, the boundaries of each sample tree are generated as separate polygons, and point cloud data is extracted from the CHM for subsequent analysis. To evaluate and optimize the automatically generated individual tree segmentation results through the MCWS algorithm, a method combining automatic segmentation with manual adjustment was adopted. After generating individual tree polygons through the MCWS algorithm, the shape file representing the polygon boundaries was imported into ArcGIS Pro, and the polygons were manually corrected using point cloud data visualization and normalized CHM. This step generated two separate sets of point cloud data for each tree: one based on automatic segmentation and another based on manually adjusted results.

Metrics Generation and Screening

After successfully extracting individual tree point cloud data, the next crucial step is to generate metrics describing the characteristics of each tree. These metrics will serve as the foundation for subsequent model construction and training. This study used the metrics_set3 function from the LidR package, which is a predefined comprehensive set of indicators capable of extracting rich tree structure information from point cloud data. Before using these metrics for model construction, pre-screening is necessary. This step aims to improve data quality and ensure the validity and reliability of model inputs. According to Bennett's (2001) study, if a variable has more than 10% of data missing, it may lead to significant bias in statistical analysis results. Based on this principle, metrics with more than 10% of data being 0 or NA and metrics without discriminative ability (i.e., metrics with the same value for all trees) were removed. Through this pre-screening process, 16 metrics were eliminated, leaving 95 effective metrics.

Model Construction

In this study, a random seed was used to allocate 70% of the 284 valid sample trees for model training, with the remaining 30% used for testing the model's predictive ability. Using a random seed ensures model reproducibility and applies the same random results to different model constructions. RStudio itself has the capability to build linear regression models, while for Random Forest, the RandomForest package is required.

This research used a fixed random seed (set.seed(123)) to randomly divide the 284 valid sample trees into a training set (70%) and a test set (30%) to ensure the reliability and reproducibility of the results. This method not only guarantees the reproducibility of the experiment but also allows for comparison between different models. For model construction, R's built-in lm function was used to build linear regression models, and the randomForest package was used to implement random forest models. Model evaluation used test set data, measuring model performance by calculating the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Among these, R^2 represents the overall fit of the model, used to evaluate the model's explanatory power for test data. R^2 ranges from 0 to 1, where $R^2 \geqslant 0.70$: Very strong correlation ; $0.50 \leqslant R^2 < 0.70$: Strong correlation ; $0.30 \leqslant R^2 < 0.50$: Moderate correlation ; $R^2 < 0.30$: Weak correlation.

The RMSE was calculated to quantify the model's prediction accuracy, and the MAE was used to evaluate the average magnitude of prediction errors.

Height Model Prediction (SLR)

Linear regression models include Single Linear Regression (SLR) and Multiple Linear Regression (MLR). The main difference between these two regression models is the number of independent variables (metrics) input. For tree height linear regression prediction models, using SLR can yield simpler and more accurate results. This is because some highly correlated indicators have a very strong linear relationship with tree height. This strong correlation means that a single LiDAR height indicator is sufficient to accurately predict tree height. The metrics include many height-related metrics (such as maximum height or 99th percentile of height), and these metrics are highly correlated with each other, which would lead to multicollinearity problems in MLR models. Multicollinearity refers to the phenomenon where two or more independent variables are highly correlated, which can affect the stability and interpretability of the model.

DBH and Volume Prediction Models (MLR)

For DBH and volume, MLR can achieve better results because DBH and volume prediction models do not rely on a single type of metric, but need to assign weights to different types of metrics to seek more accurate prediction results. Among the 95 metrics, the impact of each metric on the target variable (DBH and volume) is unknown. Therefore, it is necessary to first construct a model containing 95 metrics, and then use backward elimination to gradually remove the metrics with the lowest correlation to improve the model's interpretability and practicality while avoiding overfitting. Overfitting refers to the situation where a model achieves quite high fitting on training data but performs poorly in testing. This study mainly uses two indicators to judge the removed metrics: whether the p-value of the metric and the prediction target variable is greater than 0.05, and whether the variance inflation factor (VIF) value of the metric is greater than 5. If the p-value of a metric is greater than 0.05, it means that there is little correlation between this metric and the prediction target variable. If the VIF value of a metric is greater than 5, it indicates that this metric is highly correlated with other metrics in the model. After parameter tuning, MLR allows returning a table containing the metrics that affect the prediction target variable and their degree of influence, thereby obtaining a mathematical equation for predicting the dependent variable.

Diagnosis and Validation of MLR Models

After the MLR model is adjusted, it needs to be diagnosed. This study mainly diagnoses the model from two aspects. The first aspect is to observe whether there is a systematic relationship between adjacent residuals in the regression model by using the Durbin-Watson (DW) test with the dwtest function from the lmtest package. The DW statistic ranges from 0 to 4, where if DW is approximately 2, it indicates no autocorrelation in the residuals; when DW is less than 2, it indicates positive autocorrelation; when DW is greater than 2, it indicates negative autocorrelation. The second aspect is to conduct a comprehensive diagnosis of the linear regression model's quality using the performance, see, and patchwork packages. The diagnosis contents include the fitting degree between the model's true values and predicted values, the magnitude of residuals, whether there is a trend in residuals, whether the distribution of residual values is within an acceptable range, the collinearity of independent variables, and whether the residuals follow a normal distribution.

Machine Learning

Random Forest (RF) as an advanced machine learning algorithm has shown significant advantages in handling complex non-linear relationships, especially when analyzing LiDAR data. This ensemble learning method, by constructing and synthesizing results from multiple decision trees, not only possesses powerful predictive capabilities but also can avoid overfitting problems. Compared to MLR models, RF can not only handle non-linear relationships but also doesn't require complex parameter tuning processes. In practical applications, the implementation of RF models is relatively simple and can be achieved using the randomForest package in RStudio.

Result

Tree Height Prediction Model

After constructing and comparing SLR models and RF models using height-related metrics and tree height, the prediction model accuracy for height is generally high ($R^2 > 0.7$), indicating that LiDAR data has significant advantages in tree height estimation. The comparison between predicted height and measured height for each model is shown in Figure 4. Through horizontal comparison of all results, it was found that in the SLR height prediction model based on automatically generated individual tree segmentation, the model constructed with maximum height (zmax) as the independent variable outperformed the model constructed with the 99th percentile of height (zq99) and the two RF models generated based on different segmentation results. Its coefficient of determination R^2 reached 0.82, while the root mean square error (RMSE) was the lowest at 0.76m. This means that the average prediction error of this model is about 0.76m, demonstrating satisfactory prediction accuracy. In comparison, although the RF model can integrate all 95 metrics for comprehensive prediction, its prediction accuracy is not as good as the SLR model ($\Delta R^2 \approx 0.04$).

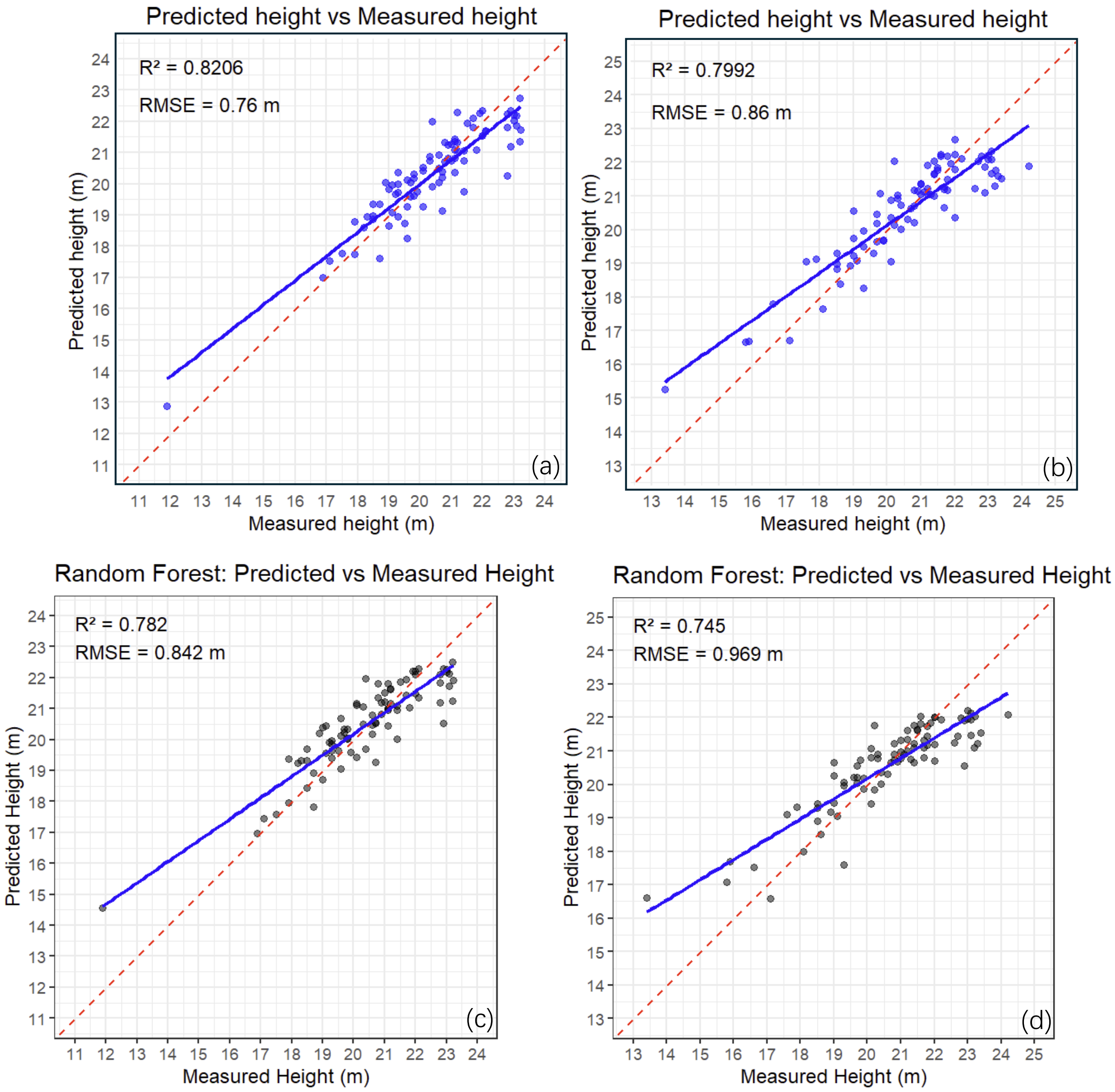


Figure 4. Comparison of tree height prediction models using different segmentation methods and modeling approaches. (a) SLR model with auto-generated segmentation (b) SLR model with manually corrected segmentation (c) RF model with auto-generated segmentation (d) RF model with manually corrected segmentation

Comparing the two sets of models generated from automatic segmentation and manually adjusted segmentation, it can be observed that the models generated from automatic individual tree segmentation significantly outperform those from manually adjusted individual tree segmentation in tree height prediction ($\Delta R^2 \approx 0.03$). This indicates that when predicting tree height, the results generated from automatic individual tree segmentation are sufficiently accurate, and there is no need for time-consuming manual corrections.

DBH Prediction Model

For DBH prediction, MLR models were used as parametric models for prediction and compared with machine learning models. Figure 5 shows the relationship between predicted DBH and measured DBH for 4 models. The DBH prediction results indicate that there is a significant difference between models generated from automatic segmentation and manually adjusted individual tree segmentation ($\Delta R^2 \approx 0.15$). The main reason is the classification errors of individual tree point clouds caused by automatic individual tree segmentation, while manual correction of segmentation fixed most of these errors, thus significantly improving the model's fit.

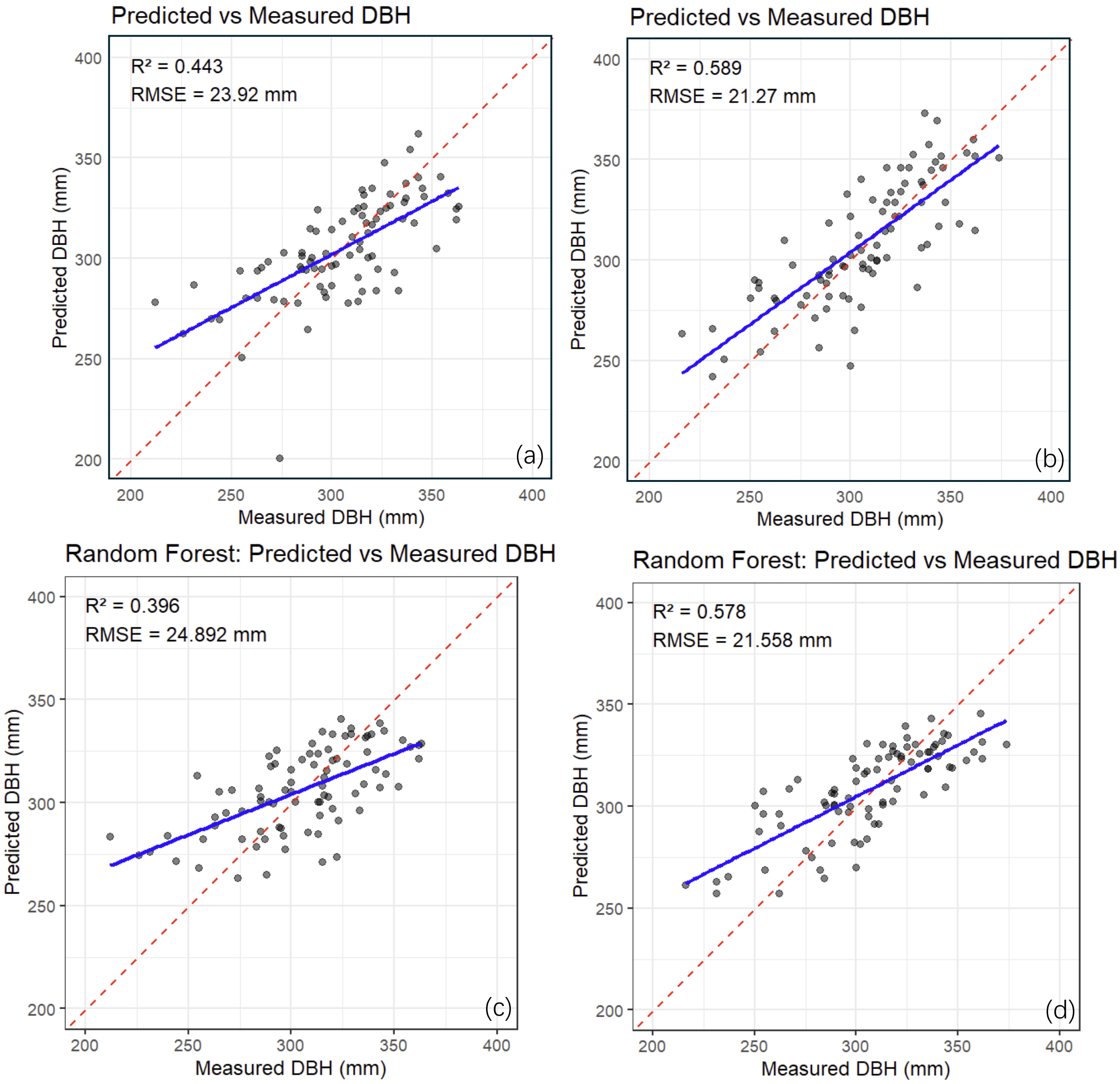


Figure 5. Comparison of DBH prediction models using different segmentation methods and modeling approaches. (a) SLR model with auto-generated segmentation (b) SLR model with manually corrected segmentation (c) RF model with auto-generated segmentation (d) RF model with manually corrected segmentation

Although predicting DBH using the predefined comprehensive set of indicators can explain the main variations affecting DBH ($R^2 > 0.5$), there is still room for improvement in the model's prediction accuracy. Similar to the comparison results of tree height prediction models, MLR models slightly outperform RF models in DBH prediction accuracy, but the difference is not significant. For example, in the two models generated after manually adjusting individual tree segmentation, the performance of MLR and RF is almost equivalent (ΔR² ≈ 0.01).

Volume Prediction Model

Consistent with the choice of DBH prediction models, MLR and RF models were used for volume prediction, with the results shown in Figure 6. Similar to the results of the DBH prediction models, there is a significant difference between models generated from automatic segmentation and manually adjusted individual tree segmentation (ΔR² ≈ 0.13). This consistency highlights the crucial role of high-quality data preprocessing (especially individual tree segmentation) when using LiDAR data for forest parameter estimation.

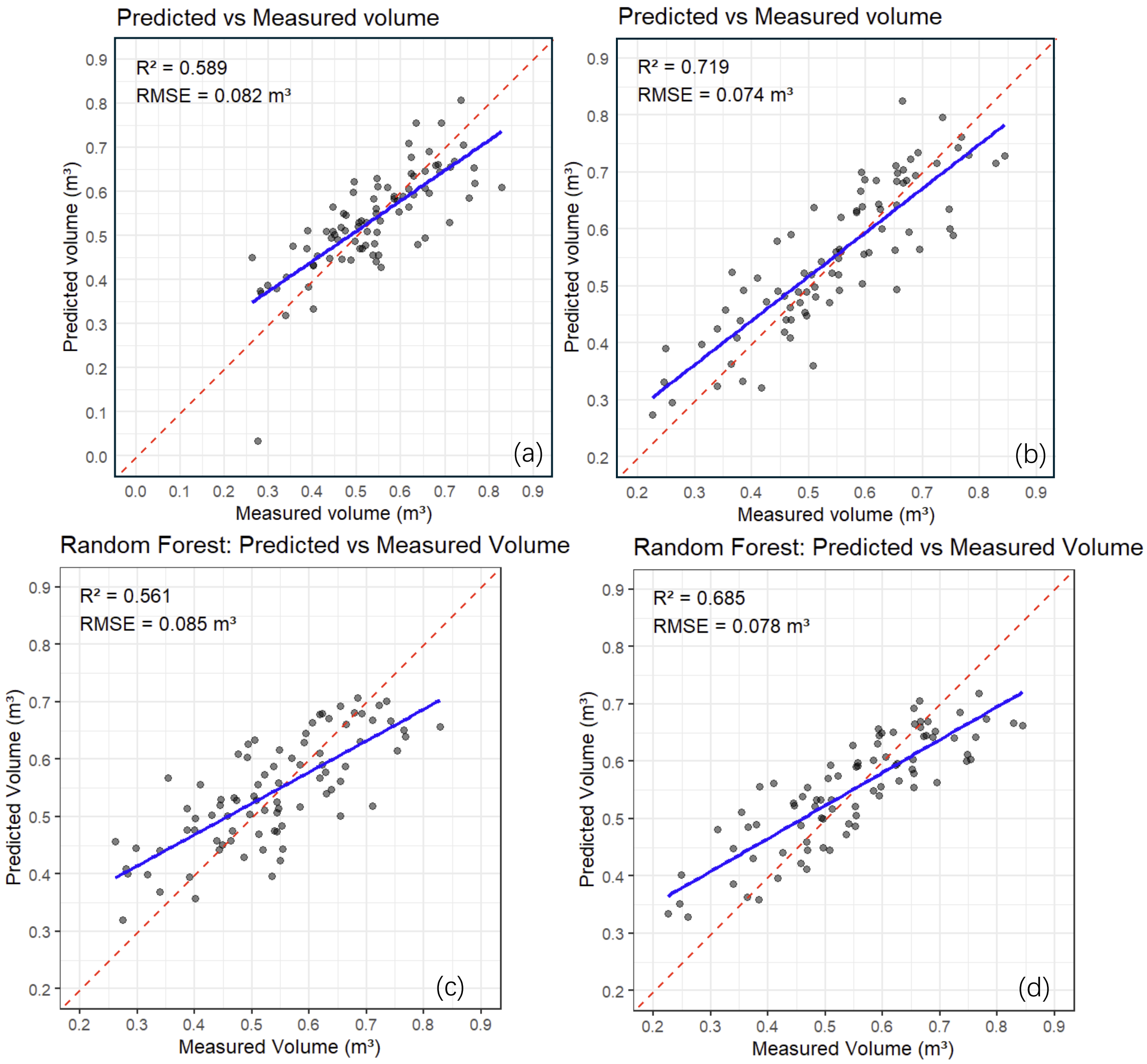


Figure 6. Comparison of volume prediction models using different segmentation methods and modeling approaches. (a) SLR model with auto-generated segmentation (b) SLR model with manually corrected segmentation (c) RF model with auto-generated segmentation (d) RF model with manually corrected segmentation

By comparing the two types of models with the same segmentation results, similar to the comparison results of DBH prediction models, the prediction accuracy of MLR models is higher than that of RF models, but the difference is not very significant ($\Delta R^2 \approx 0.03$).

Discussion

Tree height Prediction Model Results Discussion

In this project, the SLR model generated based on automatic individual tree segmentation results performed best in tree height prediction ($R^2 = 0.82$). The tree height prediction equation generated by the model is:

$$H_{predicted} = 0.9114 \times Z_{max} + 1.9304$$

Where $H_{predicted}$ is the predicted tree height, and z_{max} is the maximum height in the individual tree point cloud information.

The comparison results of four tree height prediction models for individual trees indicate that the SLR model's prediction accuracy is higher than that of the RF model. This phenomenon may be attributed to the RF model allocating weights among multiple metrics, thus diluting the weight proportion assigned to height-directly related metrics. The SLR model performs better in this case mainly because it focuses on a single metric highly related to tree height, capable of capturing tree height changes more directly. This result demonstrates that in specific situations, simple models may outperform complex ones, especially when there is a strong linear relationship between the predictor variable and the target variable.

Furthermore, the results of individual tree height prediction models show that models generated from automatic individual tree segmentation significantly outperform those from manually adjusted individual tree segmentation in tree height prediction. This result highlights the advantage of LiDAR data in capturing treetop structures, accurately reflecting tree height information even when automatic segmentation might not be precise enough. This may be because tree height is mainly determined by the highest points in the point cloud, which are usually well preserved in automatic segmentation. In contrast, manually corrected individual tree segmentation results may lead to these points being incorrectly assigned to other trees, thereby reducing the predictive ability of models generated from manually adjusted individual tree segmentation for tree height.

However, the absolute accuracy of height prediction models should be viewed with caution. This study uses manually measured tree heights as standard values and predicted values obtained from LiDAR data analysis as comparison values. This approach has potential issues: manual measurements using vertex hypsometers and performed by multiple people may lead to inconsistent measurement standards and visual errors, among other human errors. In comparison, LiDAR data undergoes multiple rounds of noise removal and compares different algorithms to obtain more accurate results. Therefore, using manually measured tree heights as a validation standard may not fully indicate the absolute accuracy of model prediction results, but rather demonstrates the potential of LiDAR data to replace traditional manual measurements in tree height prediction.

DBH Prediction Model Results Discussion

Analysis of DBH prediction results based on four different models shows that the prediction accuracy for DBH is generally low, with the highest coefficient of determination (R^2) reaching only 0.589. The prediction model results are shown in Table 2. Among them, zq90 (90th percentile of height) reflects the upper crown structure; zpcum9 (9th value of cumulative height percentage) provides information about vertical structure distribution; L3 (3rd value of L-moments statistics) describes the skewness of height distribution; lad_cv (coefficient of variation of Leaf Area Density) characterizes the spatial variability of leaf area density; pz_10.20 (percentage of points between 10-20 meters in height) reflects the point cloud density at specific height layers; p_intermidiate (proportion of intermediate returns) indicates the degree of laser penetration through the crown; vn (number of voxels after voxelization) represents the spatial distribution density of the point cloud; vzsd (standard deviation of height after voxelization) describes the degree of height variation; and coords.x1 (position information of the point cloud on the X-axis) provides spatial location data.

Table 2. Multiple Linear Regression Model Results for DBH Prediction Using LiDAR-Derived Metrics

	Estimate	Std. Error	t value	P	VIF
Intercept	-176543.021	61355.74	-2.877	0.004	
zq90	7.558	1.359	5.561	<0.001	1.773
zpcum9	1.18	0.446	2.646	0.009	1.255
L3	-31.204	11.636	-2.682	0.008	1.481
lad_cv	-31.022	10.301	-3.011	0.003	2.293
pz_10.20	-1.428	0.265	-5.392	<0.001	2.262
p_intermidiate	-8.503	1.336	-6.367	<0.001	1.929
vn	0.363	0.03	12.034	<0.001	1.469
vzsd	130.812	65.913	1.985	0.049	1.993
coords.x1	0.114	0.04	2.88	0.004	1.224
Adjusted R-squared			0.6881		
MAE%			5.55%		
RMSE			21.27 mm		
R-squared			0.5888		
WD			2.1998		

From Table 2, the expression for predicting DBH can be obtained:

$$DBH = -176543.021 + 7.558 \times zq90 + 1.18 \times zpcum9 - 31.204 \times L3 - 31.022 \times lad_{cv} - 1.428 \times pz_{10.20} - 8.503 \times p_{intermidiate} + 0.363 \times vn + 130.812 \times vzsd + 0.114 \times coords.x1$$

DBH Prediction Model Diagnostic Results

Table 2 also shows the weight and influence of each metric on the model and its VIF value. All metrics in this model have VIF values less than 5, indicating no multicollinearity. Through the DW test on the model, the model's DW is approximately 2.2, indicating a slight negative autocorrelation in the residuals, which does not seriously affect the overall performance of the model. Figure 7 further presents the diagnostic results of the multiple linear regression (MLR) model, validating the model's appropriateness and reliability by evaluating three key aspects: linearity, homoscedasticity, and normality. These diagnostic plots indicate that the DBH prediction model largely satisfies the basic assumptions of multiple linear regression. There are slight deviations in linearity and homoscedasticity, but they are not severe enough to significantly affect the overall validity of the model. The slight deviation in linearity may suggest subtle non-linear relationships between some predictor variables and DBH, while the small fluctuations in homoscedasticity may reflect natural variability in forest structure across different DBH ranges. Considering the VIF values, Durbin-Watson test results, and the analysis of these diagnostic plots comprehensively, it can be concluded that the model results are relatively reliable and can truly reflect the relationship between LiDAR-derived indicators and DBH.

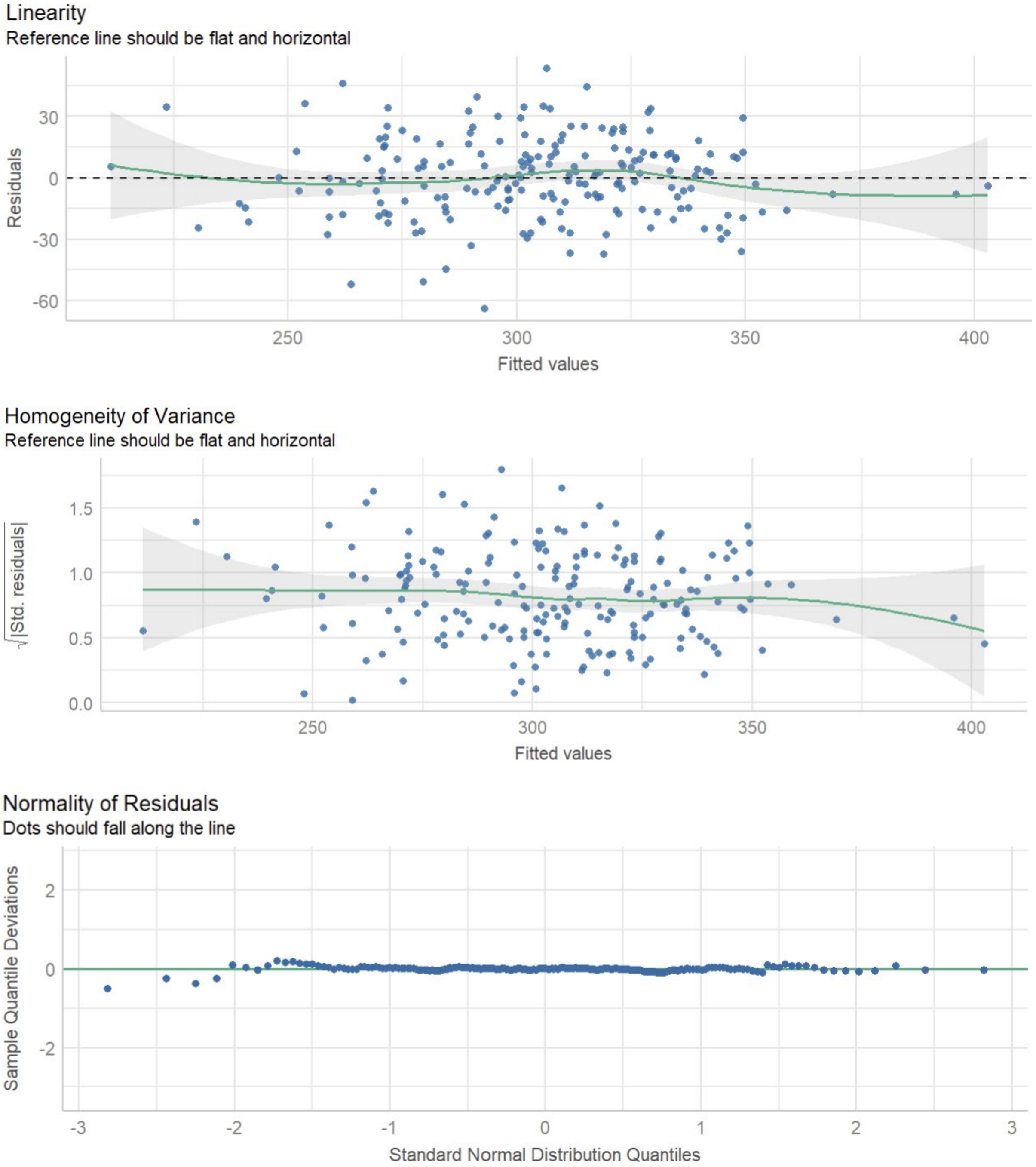


Figure 7. Diagnostic Plots for the DBH Prediction Model using Multiple Linear Regression. From top to bottom are linearity, homogeneity of variance and normality of residuals.

From the perspective of individual tree segmentation methods, the MLR model based on manually adjusted individual tree segmentation results has the highest accuracy. This indicates that the linear model can capture most of the important variation information, and the even-aged, uniform density stand structure may lead to relatively consistent tree growth characteristics, thus reducing the occurrence of complex non-linear relationships. These environmental characteristics may limit the potential advantages of more complex algorithms such as RF models, which typically perform better in handling highly non-linear and complex interactive relationships.

Compared to previous studies, the DBH prediction accuracy (R^2) in this report is lower than the range of 0.68 to 0.89 reported in the literature (Liu et al., 2018; Cao et al., 2016; Yu et al., 2011; Dalla et al., 2020). This result suggests that based on current UAV LiDAR technology and data processing methods, the accuracy of DBH estimation is not yet sufficient to fully replace traditional manual measurements. The main reason for this result is that UAV LiDAR primarily emits laser pulses from above, with most signals being obstructed by tree crowns, resulting in relatively sparse point cloud data that directly characterizes trunk features. Additionally, the predefined comprehensive set of indicators in the LidR package used in this study may not have adequately included the most critical features for DBH prediction, thus limiting the model's prediction accuracy for DBH.

Volume Prediction Model Results Discussion

Analysis of prediction results based on four different models shows that the performance of volume prediction models is significantly better than that of DBH prediction models. The highest coefficient of determination (R^2) for volume prediction reached 0.72, indicating that the model explained 72% of volume variation, reaching a strong correlation level. This result contrasts sharply with the highest R^2 (0.589) of the DBH prediction model. Detailed results of the prediction model are shown in Table 3. Notably, the LiDAR metrics used to construct the multiple linear regression (MLR) model for volume prediction are almost identical to those used in the DBH prediction model. These indicators demonstrated significantly higher accuracy in predicting tree volume, suggesting that compared to describing individual tree DBH, these metrics exhibit stronger explanatory power and predictive ability in characterizing individual tree volume features. Among them, zq90 (90th percentile of height) reflects the upper crown structure; zpcum9 (9th value of cumulative height percentage) provides information about vertical structure distribution; L3 (3rd value of L-moments statistics) describes the skewness of height distribution; lad_cv (coefficient of variation of Leaf Area Density) characterizes the spatial variability of leaf area density; pz_10.20 (percentage of points between 10-20 meters in height) reflects the point cloud density at specific height layers; p_intermidiate (proportion of intermediate returns) indicates the degree of laser penetration through the crown; vn (number of voxels after voxelization) represents the spatial distribution density of the point cloud; and coords.x1 (position information of the point cloud on the X-axis) provides spatial location data.

Table 3. Multiple Linear Regression Model Results for Volume Prediction Using LiDAR-Derived Metrics

	Estimate	Std. Error	t value	P	VIF
Intercept	-885.780793	208.784366	-4.242	<0.001	
zq90	0.046711	0.004543	10.282	<0.001	1.581
zpcum9	0.004076	0.001545	2.638	0.009	1.201
L3	-0.09142	0.039415	-2.319	0.021	1.356
pz_10.20	-0.004446	0.000867	-5.125	<0.001	1.937
p_intermidiate	-0.020711	0.004159	-4.979	<0.001	1.493
vn	0.00139	0.000106	13.062	<0.001	1.462
coords.x1	0.000572	0.000135	4.24	<0.001	1.131
Adjusted R-squared				0.7595	
MAE%				11.72%	
RMSE (m³)				0.0737	
R-squared				0.7191	
DW				2.2324	

From Table 3, the expression for predicting volume can be obtained:

$$\begin{aligned} DBH = \; & -176543.021 + 7.558 \times zq90 + 1.18 \times zpcum9 - 31.204 \times L3 - 31.022 \times lad_{cv} \\ & - 1.428 \times pz_{10.20} - 8.503 \times p_{intermidiate} + 0.363 \times vn + 130.812 \times vzsd \\ & + 0.114 \times coords.x1 \end{aligned}$$

Volume Prediction Model Diagnostic Results

Table 3 also shows the weight and influence of each metric on the model and its VIF value. All metrics in this model have VIF values less than 5, indicating no multicollinearity. Through the DW test on the model, the model's DW is approximately 2.2, indicating a slight negative autocorrelation in the residuals, which does not seriously affect the overall performance of the model. Figure 8 further presents the diagnostic results of the multiple linear regression (MLR) model, validating the model's appropriateness and reliability by evaluating three key aspects: linearity, homoscedasticity, and normality. These diagnostic plots indicate that the volume prediction model largely satisfies the basic assumptions of multiple linear regression. There are slight deviations in linearity and homoscedasticity, but they are not severe enough to significantly affect the overall validity of the model. The slight deviation in linearity may suggest subtle non-linear relationships between some predictor variables and volume, while the small fluctuations in homoscedasticity may reflect natural variability in forest structure across different volume ranges. Considering the VIF values, Durbin-Watson test results, and the analysis of these diagnostic plots comprehensively, it can be concluded that the model results are relatively reliable and can truly reflect the relationship between LiDAR-derived indicators and tree volume.

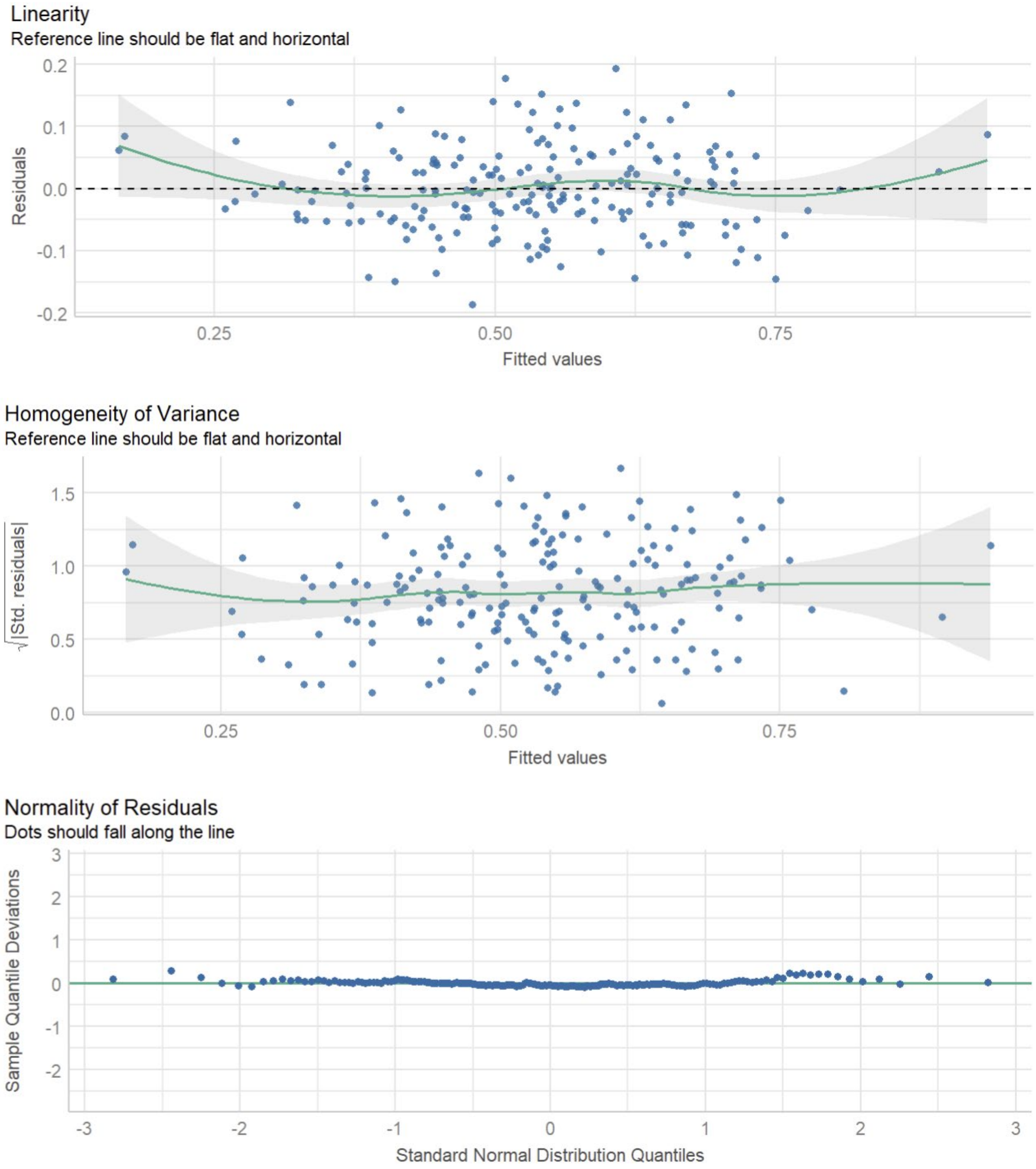


Figure 8. Diagnostic Plots for the Volume Prediction Model using Multiple Linear Regression. From top to bottom are linearity, homogeneity of variance and normality of residuals.

Compared to previous studies, the R² of the volume prediction model falls within the range of 0.7 to 0.93 reported in the literature (Liu et al., 2018; Yu et al., 2011; Hayashi et al., 2014), indicating that the metrics used to construct the model can well describe about 72% of volume variation. The predefined comprehensive set of indicators used in this study already includes features that are relatively critical for volume prediction, thus making the model's prediction accuracy for volume reliable.

Compared to the diagnostic results of the DBH prediction model, the volume prediction model performs better in meeting regression assumptions. The volume model demonstrates higher prediction stability and resistance to outliers, while the DBH model shows stronger non-linear relationships and sensitivity to outliers. This difference may stem from a more direct relationship between tree volume and the three-dimensional structural features captured by LiDAR, while DBH, as a two-dimensional measurement, is difficult to capture directly in LiDAR data obtained from above. The random forest model's performance approaching that of multiple linear regression in volume prediction suggests that there may be some non-linear relationships between volume and LiDAR indicators, although these non-linear features may not be as significant as expected.

When using the traditionally manually measured DBH and tree height, and the volume estimated through formulas for individual trees as standard values, the prediction model can achieve a strong correlation. Consistent with the issues implied by tree height measurements, the volume estimation formula based on DBH and tree height as input parameters may lead to significant manual errors in volume estimation. Therefore, using the predicted values obtained from LiDAR data analysis as a comparison can only prove the feasibility of replacing traditional volume estimation formulas with UAV LiDAR, but cannot prove whether higher prediction accuracy can be achieved through UAV LiDAR.

Future Study

Further research can focus on improving DBH prediction accuracy and optimizing automatic segmentation algorithms. Given the complex relationship between DBH and LiDAR-derived indicators, more refined modeling strategies should be considered, such as exploring advanced machine learning algorithms, designing specialized DBH-related LiDAR indicators, and studying the association between DBH and crown structure characteristics. Although automatic segmentation based on the MCW algorithm performs well in tree height prediction, DBH and volume prediction still require manual verification to improve accuracy. Improving automatic segmentation algorithms can start from aspects such as optimizing MCW parameters, applying deep learning techniques, and researching adaptive segmentation algorithms. These optimizations can not only avoid time-consuming manual corrections but also improve the overall efficiency and accuracy of LiDAR data analysis.

Conclusion

This study constructed models for predicting tree height, diameter at breast height (DBH), and volume by analyzing UAV LiDAR point cloud data and extracting key metrics, while comparing automatic individual tree segmentation based on the MCW algorithm with manually corrected segmentation results. A total of 12 prediction models were constructed, with the best fit (R^2) reaching 0.82 for tree height, 0.59 for DBH, and 0.72 for volume, with linear models consistently outperforming random forest models. The study found that automatic segmentation performed excellently in tree height prediction, but showed significant differences from manually corrected results in DBH and volume prediction. These results confirm that UAV LiDAR-based prediction models can replace traditional labor-intensive measurement methods in tree height and volume estimation, improving the efficiency of forest surveys. However, the relatively lower accuracy of DBH prediction indicates that challenges remain in estimating this parameter. Therefore, future research should focus on improving DBH prediction accuracy and optimizing automatic segmentation algorithms to further enhance the overall prediction accuracy and practicality of the models. These improvements will provide stronger support and valuable data for areas such as precision forestry, sustainable forest management, and large-scale forest monitoring.

References

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25*(5), 464-469. https://doi.org/https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Cao, L., Gao, S., Li, P., Yun, T., Shen, X., & Ruan, H. (2016). Aboveground biomass estimation of individual trees in a coastal planted forest using full-waveform airborne laser scanning data. *Remote Sensing, 8*(9), 729. https://doi.org/10.3390/rs8090729

Coomes, D. A., Dalponte, M., Jucker, T., Asner, G. P., Banin, L. F., Burslem, D. F. R. P., Lewis, S. L., Nilus, R., Phillips, O. L., Phua, M.-H., & Qie, L. (2017). Area-based vs tree-centric approaches to mapping forest carbon in Southeast Asian forests from airborne laser scanning data. *Remote Sensing of Environment, 194*, 77-88. https://doi.org/10.1016/j.rse.2017.03.017

Dalla Corte, A. P., Rex, F. E., Almeida, D. R. A. d., Sanquetta, C. R., Silva, C. A., Moura, M. M., Wilkinson, B., Zambrano, A. M. A., Cunha Neto, E. M. d., Veras, H. F. P., de Moraes, A., Klauberg, C., Liesenberg, V., Temporal, W. G., Gaiad, N. P., Zanon, M. L. B., & Broadbent, E. N. (2020). Measuring individual tree diameter and height using GatorEye high-density UAV-Lidar in an integrated crop-livestock-forest system. *Remote Sensing*, *12*(5), 863. https://doi.org/10.3390/rs12050863

Dalponte, M., & Coomes, D. A. (2016). Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data. *Methods in ecology and evolution*, *7*(10), 1236-1245. https://doi.org/10.1111/2041-210X.12575

Hayashi, R., Weiskittel, A., & Sader, S. (2014). Assessing the feasibility of low-density LiDAR for stand inventory attribute predictions in complex managed forests of northern Maine, U.S.A. *Forests*, *5*(2), 363-383. https://doi.org/10.3390/f5020363

Hyyppä, J., Yu, X., Hyyppä, H., Vastaranta, M., Holopainen, M., Kukko, A., ... & Alho, P. (2012). Advances in forest inventory using airborne laser scanning. *Remote sensing*, *4*(5), 1190-1207.

Liu, K., Shen, X., Cao, L., Wang, G., & Cao, F. (2018). Estimating forest structural attributes using UAV-LiDAR data in Ginkgo plantations. *ISPRS Journal of Photogrammetry and Remote Sensing, 146*, 465-482. https://doi.org/10.1016/j.isprsjprs.2018.10.012

Manning, A. (2023). The uptake and barriers of geospatial technologies in New Zealand's plantation forestry sector. UC Research Repository.

Mei, Z., Chun-gan, L., & Zhen, L., & Zhu, Y. (2023). Effect of UAV-LiDAR Point Density on Estimation Accuracy of Forest Inventory Attributes. *林业科学研究, 37*(2), 39-47. https://doi.org/10.12403/j.1001-1498.20230242

Michael, s., Watt, Sadeenpa, J., Robin, J., L., Hartley, Grant, D., Pearse, Peter, D., Massam, David, C., Benjamin, S., C., Steer & Honey, J., C., Estarija. (2024). Use of consumer-grade UAV laser scanner to identify trees and estimate key tree attributes across a point density range. *Forest, 15*(6), 899. https://doi.org/10.3390/f15060899

Posit team. (2023). RStudio: Integrated development environment for R. Posit Software, PBC. http://www.posit.co/

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Shugart, H. H., Asner, G. P., Fischer, R., Huth, A., Knapp, N., Le Toan, T., & Shuman, J. K. (2015). Computer and remote-sensing infrastructure to enhance large-scale testing of individual-based forest models. *Frontiers in Ecology and the Environment*, *13*(9), 503-511. https://doi.org/10.1890/140327

Torresan, C., Berton, A., Carotenuto, F., Di Gennaro, S. F., Gioli, B., Matese, A., … Wallace, L. (2016). Forestry applications of UAVs in Europe: a review. *International Journal of Remote Sensing*, *38*(8–10), 2427–2447. https://doi.org/10.1080/01431161.2016.1252477

Xu, C., Manley, B., & Morgenroth, J. (2018). Evaluation of modelling approaches in predicting forest volume and stand age for small-scale plantation forests in New Zealand with RapidEye and LiDAR. *International Journal of Applied Earth Observation and Geoinformation*, 73, 386-396. https://doi.org/10.1016/j.jag.2018.06.021

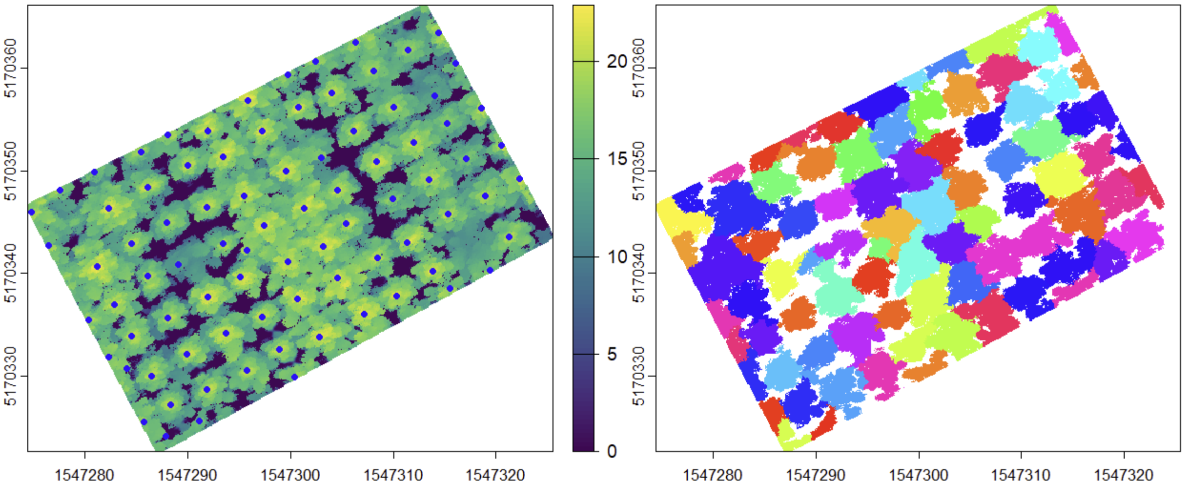
Yu, X., Hyyppä, J., Vastaranta, M., Holopainen, M., & Viitala, R. (2011). Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(1), 28-37. https://doi.org/10.1016/j.isprsjprs.2010.08.003

Appendix

Tree detection and segmentation result of grid search for each plot

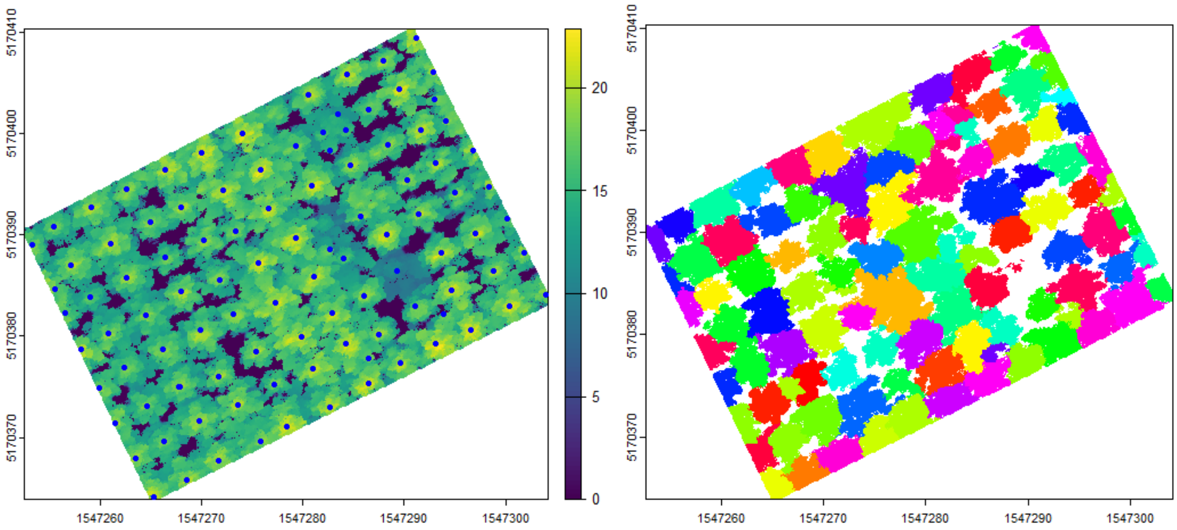
Plot 31

0.06*x+0.5



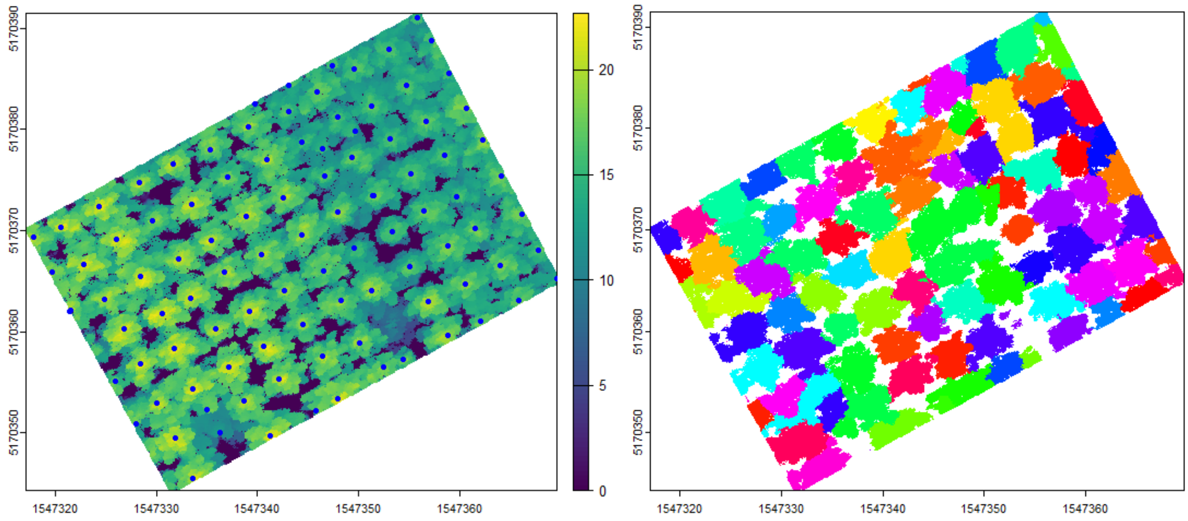
Plot 32

0.04*x+0.7



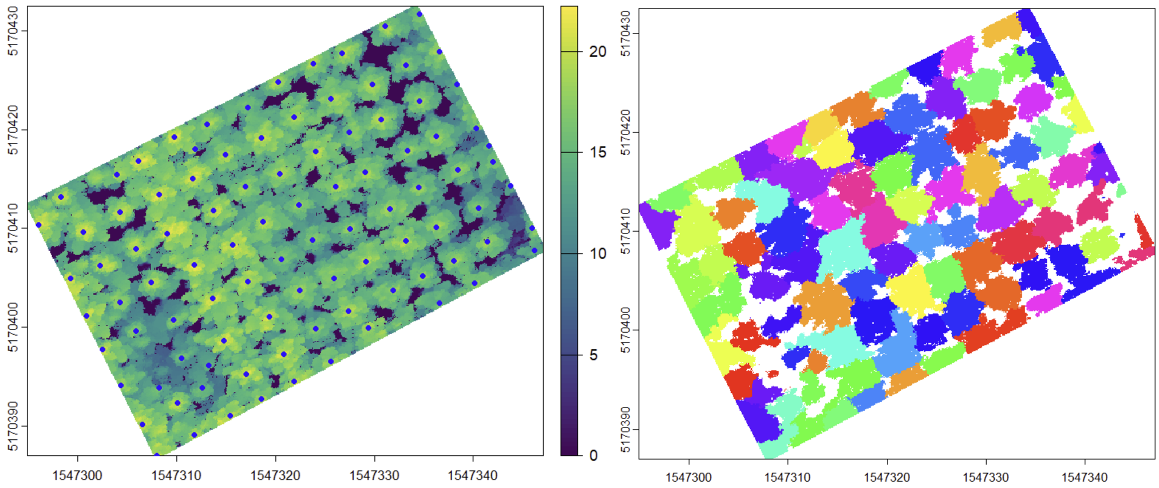
Plot 35

0.07 * x + 0.4



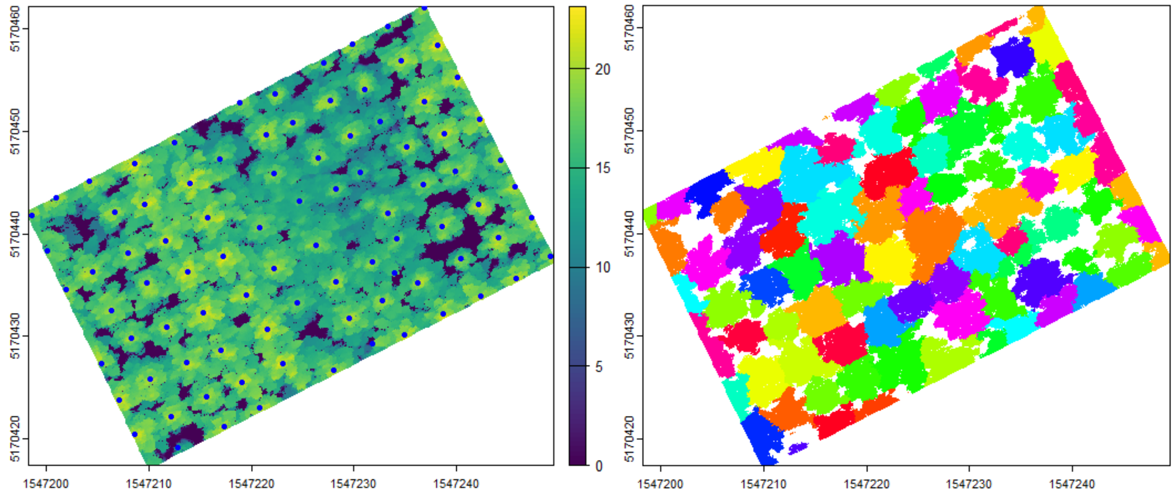
Plot 36

0.07 * x + 0.4



Plot 41

0.05 * x + 0.8


5170460
5170450
5170440
5170430
5170420
1547200
1547210
1547220
1547230
1547240
20
15
10
5
0