Forest Biomass Prediction: An Application of Multiple Linear Regression

Group Name: Hulu-Wa!

Hanwen Guan(hg2636), Yunning Liu(yl5202), Shuhan Mao(sm5322), Yi Sun(ys3594)

Runsheng Wang(rw2967), Weijie Xia(wx2281), Hongyan Zhou(hz2827)

Dec 15th, 2022

Abstract

This study seeks to predict AGBt (total biomass of all aboveground tree components) on Biomass and Its Allocation in Chinese Forest Ecosystems using multiple linear regression. Training and validation sets were first separated. Missing and inconsistent data were addressed via imputation and removal methods. Multiple transformations on the response variable were considered and log transformation was applied to stabilize variances and potentially normalize regression errors. Outliers in the response were identified and winsorized, before feature engineering methods were applied. LASSO and Random Forest were used to identify significant continuous features, while Tukey HSD was used to collapse categorical levels and identify significant levels. PCA was applied to a subset of predictors to reduce multicollinearity and reduce dimensionality. A multiple linear regression model was fitted with MAP, Stand age, Larix forest, Other Forest type, Pinus tabuliformis, and principle components selected by PCA. Diagnostic analysis using Q-Q plot, residual histogram, and residual plot does not show violations in model assumption. Training and validation were then performed and validation curves were constructed to assess the goodness of the prediction. From the analysis, a conclusion can be drawn that a MLR model with these covariates predicts AGBt quite well, but more diverse data are needed to take a comprehensive look into the Chinese forest ecosystem.

1 Introduction

Forest biomass and its allocation have long been crucial factors in forest ecosystem structure and function. They can be useful in exploring many ecological questions such as forest community dynamics, life-history evolution, and terrestrial carbon cycling. To predict biomass level, both internal factors (e.g., forest types, or soil fertility), as well as external factors (e.g., geographical location, climate, or precipitation), should be taken into consideration.

In this study, we seek to explore and identify parameters that have a significant impact on Chinese forest biomass. After preprocessing the raw dataset (imputing missing data, doing transformations) and carrying out feature selection (LASSO & PCA), we mainly focus on building a multiple linear regression model to predict AGBt (total biomass of all aboveground tree components) using 6 features: MAP, Stand Age, PCA factor, Larix Forest, Other Forest Type, and P tabuliformis. The resulting model will be tested for model assumptions and its predictive power will be measured using the validation dataset. We seek to make better predictions for Chinese forest biomass, which would contribute to sustainable development both ecologically and economically.

2 Background

The dataset contains data from 1978 to 2008 on biomass and its distributions in China's forests (excluding Hong Kong, Macao, and Taiwan). The columns of this dataset can be divided into two parts: biomass data (tree overstory components, understory vegetation, etc.) and external variables(geographical location, climate, etc.). 1607 entries for 348 study sites and broad climatic gradients (-5.1–23.8°C in mean annual temperature and 223–2515 mm in mean annual precipitation) are covered in the dataset.

3 Motivation

In this study, we decided to use AGBt, above-ground biomass total, as our response variable.

The definition of biomass is renewable organic material that comes from plants and animals. Specifically, we would use biomass as an energy source like fossil fuels for heat, electricity, and biofuel. In the meantime, the advantages of biomass are fewer air emissions, less pollution in landfills, and a decrease in our reliance on fossil fuels. Moreover, the amount of biomass is positively related to the forest condition as well. Therefore, to understand this energy and forest better, in our case, we are going to examine the most apparent part of the forest, which is all plants above ground. Fortunately, by training the data in a proper model, we can find some significant properties which could influence biomass, and biomass, also as a great indicator of forest condition, we could utilize this information to protect and maintain the ecosystem.

4 Modeling and Analysis

4.1 Train Test Split

To better assess the model's generalizability and avoid overfitting, the input dataset was randomly split into one training set containing 80% of the overall data, and one validation set with the remaining data. A random seed of 0 was set to guarantee reproducibility. The resulting split produced a training set of cardinality 1285 and a validation set of cardinality 322

4.2 Missing Values and Inconsistency

Figure 1 below shows the distribution of missing values in the dataset. An entry has a solid color if the corresponding value of the cell is not null.

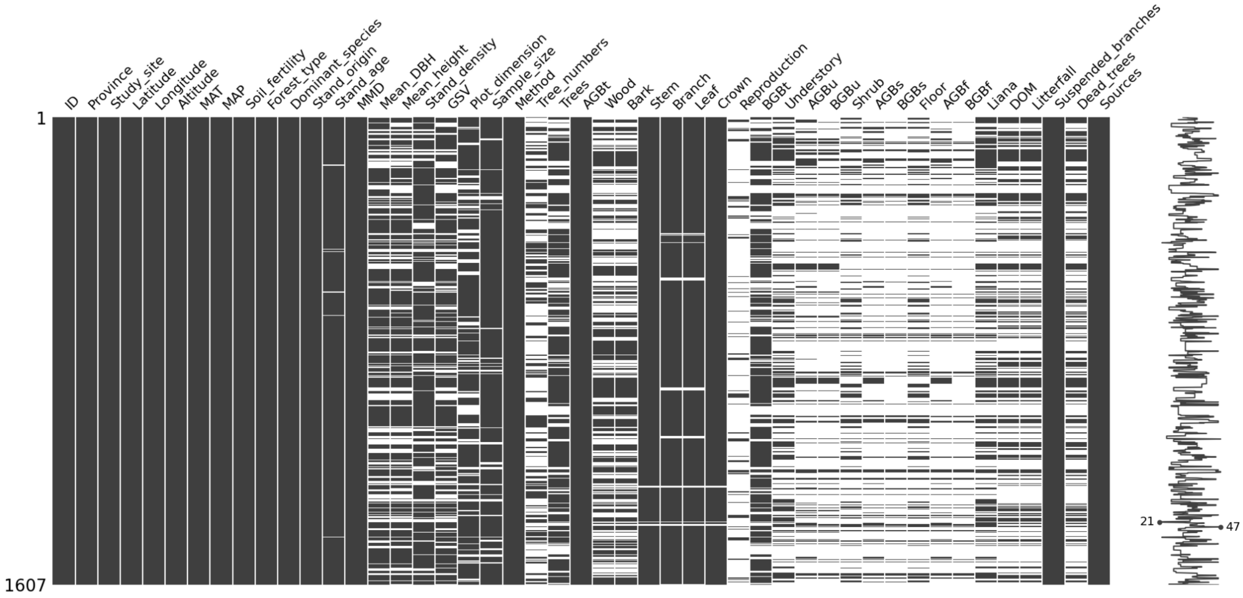


Figure 1. Missing Values Chart

As the figure suggests, some columns are composed primarily of missing values. Imputation techniques would not be suitable for these columns, as there is not enough information to infer any statistics in these columns. Specifically, columns with more than 10% of the entries as NaN were removed from the dataset. Based on the data dictionary/metadata repository, columns containing metadata that are not useful for the analysis were also removed. We also checked for duplicate data entries. Entry at index 461 was found to be a duplicate entry, which was removed.

4.3 Imputation

After addressing missing values and inconsistent data, the dataset reduces to fifteen columns, with five columns containing missing values. The maximum number of missing entries among these five columns is 46, which is approximately 3.5% of the total number of observations. We felt that an imputation on this scale is reasonable and useful, and a median imputation was applied to fill in the missing values for these columns. It is worth noting that the train test split was performed before any imputation to prevent leakage from training to testing and guarantee objectivity when evaluating validation set performances.

4.4 EDA

Figure 2 below shows the scatterplot matrix of the dataset. There are a few elements worth highlighting. First, there are potential outliers in the dataset, which can be seen from both the histogram on the diagonal and the extreme values on the individual scatterplots. Secondly, there are some collinearities among the predictors, namely: Latitude-MAT, Latitude-MAP, MAT-MAP, and Branch-Crown. Note that these collinearities are to be expected based on geological domain knowledge (e.g., higher latitude corresponds to lower temperatures). Thirdly, the biomass-related predictors (Stem, Branch, etc) appear to show some strong correlation with the response, yet they are correlated among themselves. We will demonstrate our approach to addressing these findings in the following sections. Note that standardization of the continuous predictors was applied after EDA to make their scales comparable and to make it easier to interpret the results.

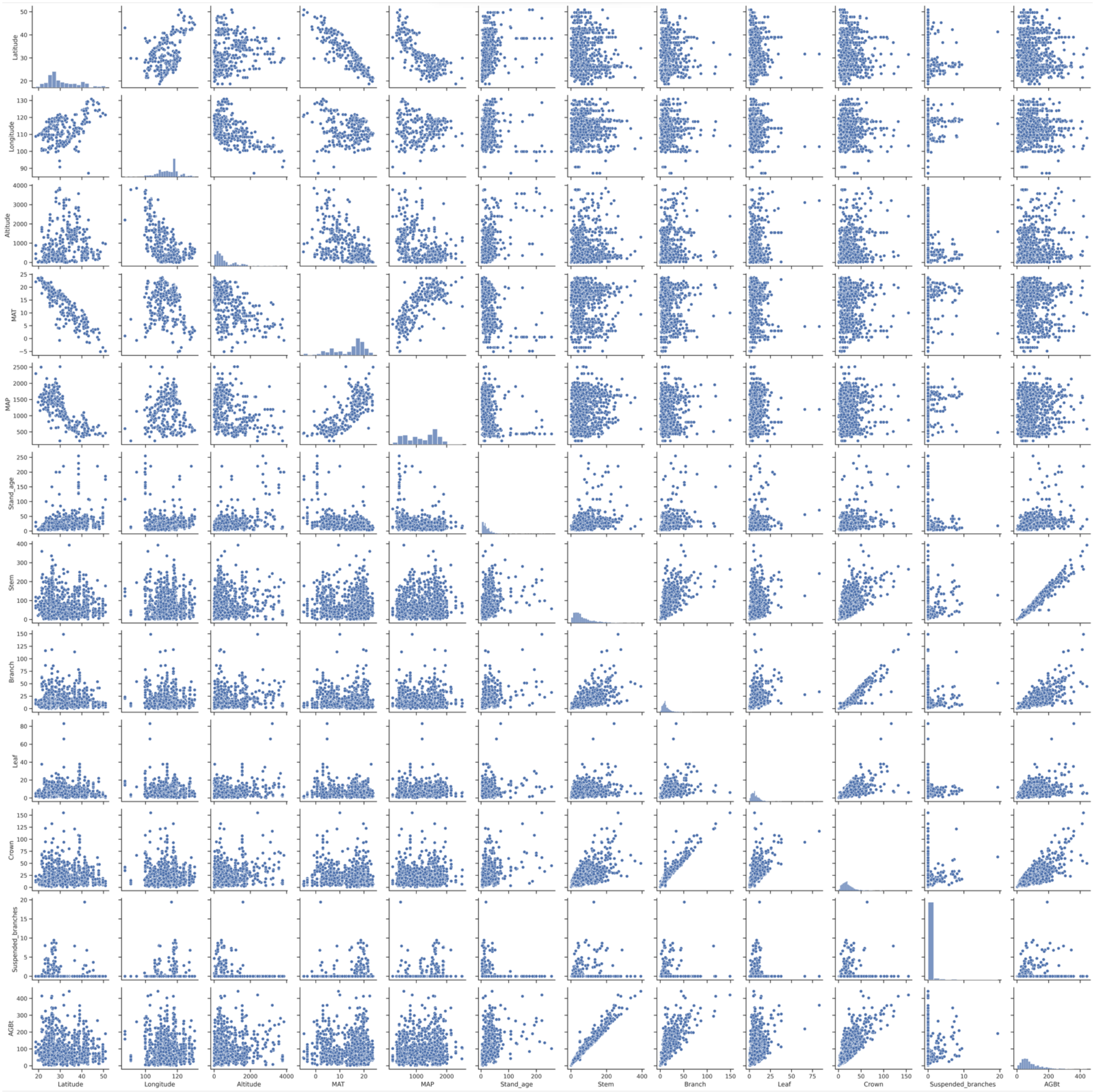


Figure 2. Scatterplot Matrix

For the categorical predictors, we computed their class distributions and noted that there is a class imbalance for two of the predictors. Predictor 'Forest_type' contains 60 unique classes, while predictor 'Dominant_species' contains 256 unique classes. Furthermore, the majority of the class labels in both of these variables account for only a very small proportion of the overall observation (e.g., 1 or 2 observations), while a few classes dominate the dataset, accounting for up to 20% of the overall observations. This result prompts us to collapse the class labels into fewer categories, which we perform later in the analysis.

4.5 Transformation

Upon initial inspection, the response variable appears to exhibit a heavy positive skew. Since errors are assumed to be i.i.d normal under the regression assumptions, transformations were considered in hope of stabilizing variances and normalizing errors. To identify the transformation to be applied, Tukey's Ladder of Power was used on the response variable. Tukey's ladder function considers a range of ƛ values for transformation and chooses the one that maximizes the Shapiro-Wilks W statistic. The resultant ƛ is the transformation that produces the most "normal-like" distribution.

$$y = \begin{Bmatrix} x^{\lambda} & \text{if } \lambda > 0 \\ \log x & \text{if } \lambda = 0 \\ -(x^{\lambda}) & \text{if } \lambda < 0 \end{Bmatrix}$$

Figure 3. Tukey's Ladder of Power

However, we hope to preserve some explainability in the response variable, and a power transformation, especially Box-Cox, is nearly impossible to interpret on its original scale, whereas log transformation can be seen as a percentage change. We decided to sacrifice some normality in exchange for the ease of interpretation of log transforms and its common usage in statistical practices. The pre-transform and post-transform histograms for the response variables are plotted below, with the colors corresponding to different bin scales.

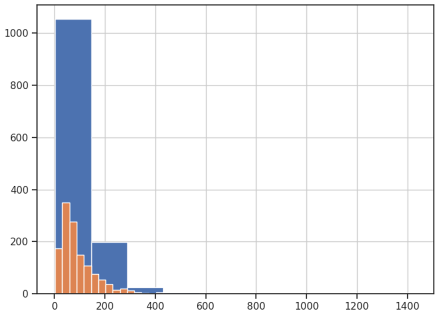


Figure 4. Histogram of Y Pre-Transform

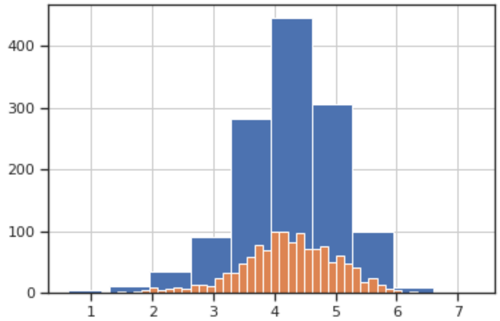


Figure 5. Histogram of Y Post-Transform

4.6 Outlier

Since forest biomass is dependent on a plenty of variables and our dataset appears to contain outliers based on the scatterplot, it is worth performing some preprocessing to address this issue. In step 3.5, the training response was transformed to be approximately normal. Therefore, bootstrapping criteria from the normal distribution to identify outliers is an appropriate method. Box plots and their whiskers extend the idea of the Empirical Rule to near-normal distributions by approximating the behaviors of Z-Scores from the normal distribution. Whiskers are usually defined to be $Q1 - 1.5(IQR)$ and $Q3 + 1.5(IQR)$. To express the whiskers in terms of σ in the normal distribution, the quantiles need to be calculated. The first and third quartiles lie at -0.675σ and +0.675σ.

$$Lower\ Bound = Q1 - 1.5(IQR) = Q1 - 1.5(Q3 - Q1) = (-0.675)\sigma - 1.5(0.675\sigma - (-0.675\sigma)) = (-2.7)\sigma$$

$$Upper\ Bound\ = Q3\ +\ 1.5(IQR)\ =\ Q3 + 1.5(Q3 - Q1) = (0.675)\sigma + 1.5(0.675\sigma - (-\ 0.675\sigma)) = (2.7)\sigma$$

The area under the normal curve outside of -2.7σ and 2.7σis approximately 0.0069, which is extremely small. Thus, points outside of the whiskers could be considered as outliers. Since outliers are still actual observations, removing them is equivalent to losing valuable information. Therefore, we chose to winsorize them by performing quantile capping and quantile flooring at the whiskers. Using this method, we found that the dataset contains 30 y-values that are beyond the whiskers, and these values were winsorized to their respective upper/lower whiskers. Figure 6 and figure 7 show the box and whiskers plot of the transformed response variable before and after winsorizing.

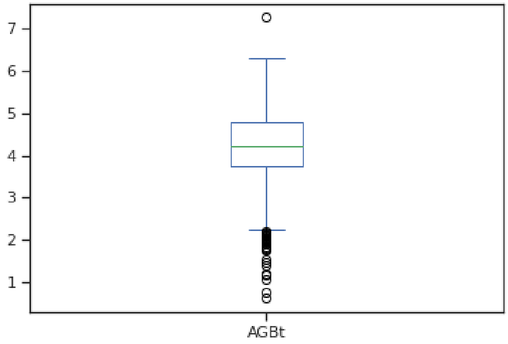


Figure 6. Boxplot Pre-Winsorizing

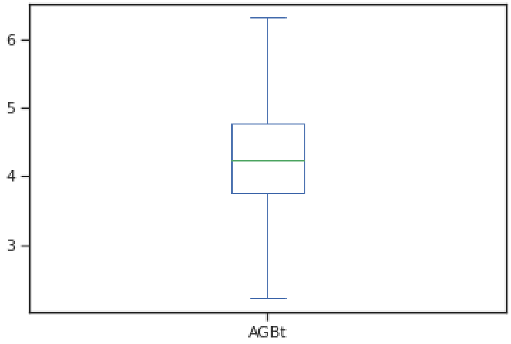


Figure 7. Boxplot Post-Winsorizing

4.7 Feature Engineering

4.7.1 LASSO

In order to make a statistically justified feature selection and obtain the best model possible, LASSO was applied to the training set. LASSO is a variation on the least square minimization method with a penalty on the L1 Norm. The penalty allows for feature selection based on λ (penalty coefficient) values and prevents overfitting by shrinking unimportant s to zero. Figure 8 demonstrates the equation used to compute LASSO.

$$\hat{\beta}_{lasso}(\lambda) = \underset{\hat{\beta}^*}{\operatorname{argmin}} \underbrace{||\vec{y} - X\hat{\beta}^*||_2^2}_{\text{loss}} + \lambda \underbrace{||\hat{\beta}^*||_1}_{\text{penalty}}$$

Figure 8. LASSO Equation

Specifically, we applied LASSO to all the continuous features except the biomass-related features (Stem, Branch, Leaf Crown). We deliberately avoided applying LASSO to the biomass-related features because we will be transforming these four features with PCA later. We also excluded all the categorical features because we will be performing Tukey HSD to identify categorical features of importance, and that LASSO on categorical features is only meaningful if grouped LASSO is performed on one-hot encoded features (which is very computationally expensive).

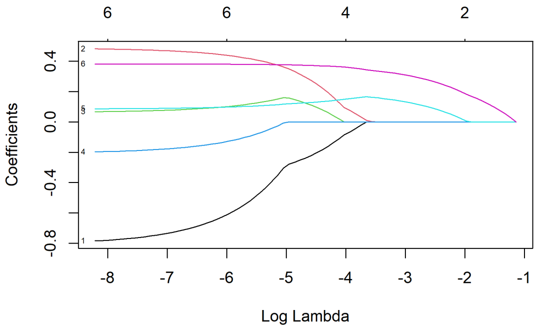


Figure 9. LASSO Shrinkage by Log λ

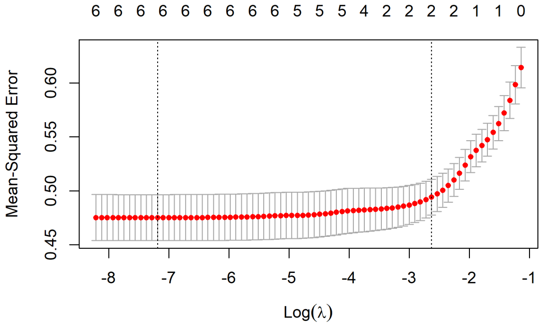


Figure 10. LASSO Cross Validation

Figure 9 shows how the coefficient of each predictor shrinks as log(λ) increases. As log(λ) increases, predictors that shrink to zero later are more significant than predictors that shrink to zero earlier. To determine the best predictors and the best number of predictors for the specific training dataset, a 10-fold cross-validation was performed on a sequence of λs. The range of λ was not specified, and glmnet chooses its own sequence to test. The cross validation function computes the fit using 10 fold cross validation for each λ in its selected range. The average error and standard deviation over the folds is then computed for each λ. Figure 10 shows the result of LASSO cross validation by plotting log(λ) against MSE with confidence intervals. Two λ values are returned: λ_{min} (left vertical line) and λ_{1se} (right vertical line). λ_{min} is the λ that gives minimum MSE among all λs. λ_{1se} gives the most regularized model such that the error is within one standard error of the minimum. Given the specific shape of the cross validation curve, λ_{1se} seems to be the better choice for two reasons. Firstly, λ_{1se} is a larger λ that corresponds to a stronger penalty, which means a more strict standard when selecting which predictors are significant. λ_{1se} produces a simpler model compared to λ_{min}, which avoids overfitting while maintaining an accuracy that's comparable to the best model. Secondly, The MSE on λ_{min} is very close to the MSE at λ_{1se}, meaning that by picking λ_{1se} we won't be sacrificing a lot of predictive power while enjoying the benefit of a regularized fit.

Based on the λ_{1se} criterion, the only two features that are still significant after the penalty are feature #5 and feature #6, which correspond to MAP and Stand_age in our dataset. All the remaining features are not considered for regression.

4.7.2 PCA

The biomass-related variables show some strong correlations with the response variable while exhibiting multicollinearity among them. Since they are all describing a similar aspect of the forest, we decided to apply PCA to avoid overfitting and reduce collinearity. PCA reduces the dimensionality of our dataset by identifying and preserving dimensions with large variances and approximating our dataset in a k-dimensional space by projecting the original data onto the reduced space. This method is better than performing feature selection and removing less important features because the information in the less important features is preserved during the projection. The resulting k-dimensional features are also linearly independent, which solves the problem of multicollinearity.

We specified k to be 2, and we noted that the top two PCA components account for 70.967577% and 18.837346% of the variations in the original dataset, respectively. The corresponding eigenvalues are 60.37296378 and 31.10443808, and the new basis vectors are [0.47019545 0.54191998 0.38682142 0.57931698] and [-0.4315073 -0.28096348 0.85624377 0.04132291]. Since we have a random seed setting, these results are fully reproducible with our code. We label the newly obtained PCA vectors as pca_0 and pca_1 in our dataset.

4.7.3 Random Forest Feature Importance

By levering the impurity criterion of the splits, Random Forest can be used as a method for feature selection. A random forest is a set of Decision Trees, where each tree contains nodes at which the dataset is being split based on one of the features. A good split is defined as when the resulting split produces subsets with similar characteristics, where the notion of 'similarity' is usually measured by the Gini impurity. Figure 11 shows the formula for computing the Gini impurity, where t is a node or a subtree, c is all the class labels, and $p(i|t)$ is the probability of a random observation in subtree t to take on a class label of i.

$$Gini(t) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

Figure 11. Gini Formula

To obtain the feature importance, we first compute how each feature decreases this impurity on all trees in the Random Forest, then we compute the average decrease with respect to each feature. The computed feature importance bar chart is shown in Figure 12.

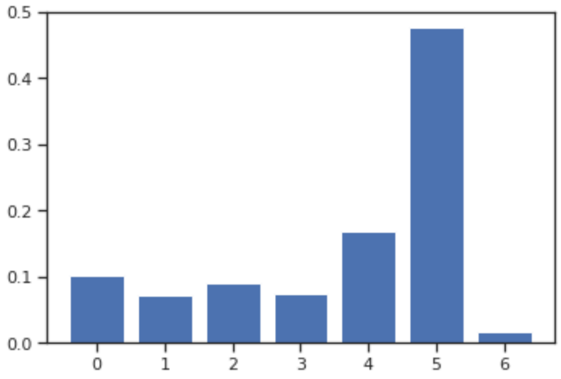


Figure 12. Random Forest Feature Importance

It is worth noting that the absolute value of the bars in Figure 12 is meaningless; feature selection should be based on the relative scale of these bars. The numerical output shows that feature #6 has an importance score of 0.01597, which is 5 times smaller than the second smallest feature importance value, while all other values appear to have similar importance scales, except for feature #5. Therefore, we choose to exclude feature #6 from our predictors, which corresponds to 'Suspended_branches' in the original dataset.

4.7.4 Log Transformation of Predictors

Before moving on to categorical feature engineering, we checked the scatterplot between the remaining continuous features and log(AGBt). We noticed that Stand_age and pca_0 exhibit an exponential relationship with log(AGBt). Therefore, we log transformed Stand_age and pca_0.

4.7.5 Tukey HSD

As described earlier in EDA, selecting categorical features from this dataset is a challenging task, primarily because two of these features have a high number of levels. To perform Tukey HSD on these features, we will first need to collapse the classes. For 'Forest_type', the top five classes in terms of frequency amount to more than half of the dataset, while for 'Dominant_species', the top five classes in terms of frequency amount to approximately half of the dataset. We decided to retain the top five classes in each variable, and encode everything else to one category called "Others". In this way, we obtain six possible classes for each of the two variables. Now we can apply Tukey HSD on all four categorical variables. The outputs are shown below from Figure 13 through 20.

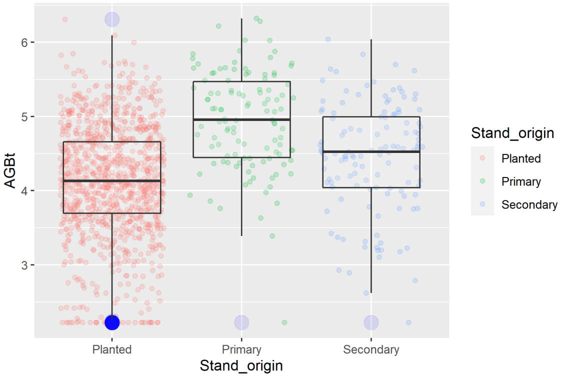


Figure 13. Box plot for Stand_origin

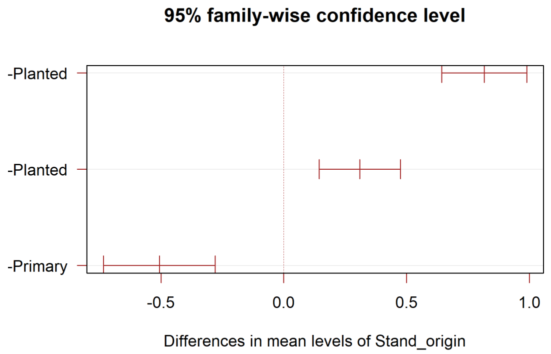


Figure 14. Tukey HSD for Stand_origin

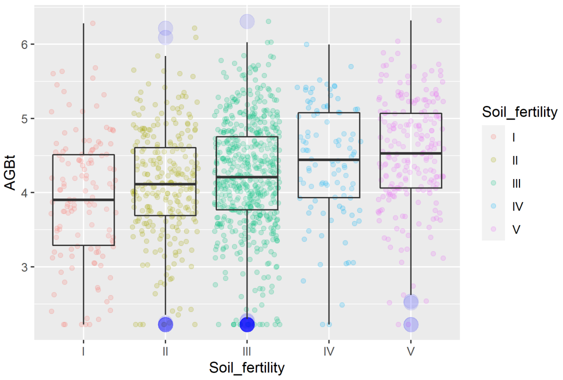


Figure 15. Box plot for Soil_fertility

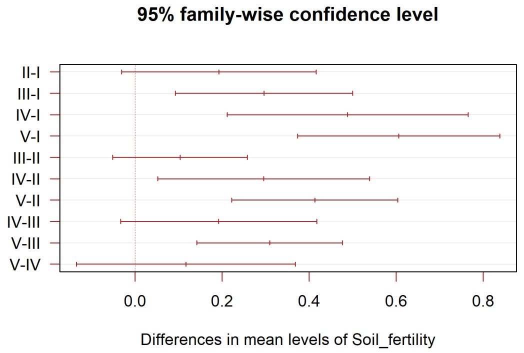
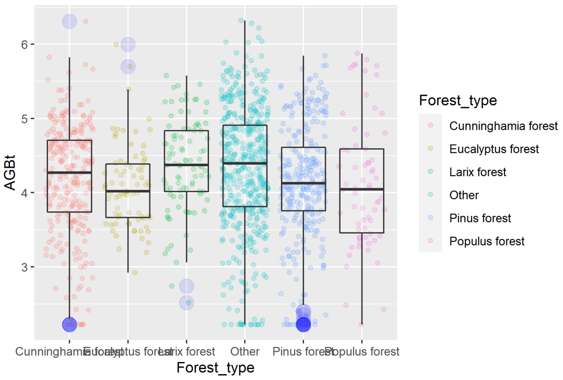


Figure 16. Tukey HSD for Soil_fertility



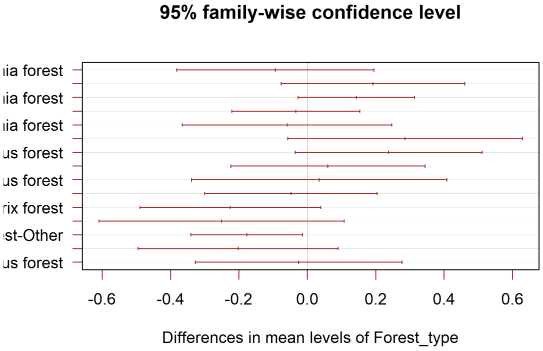
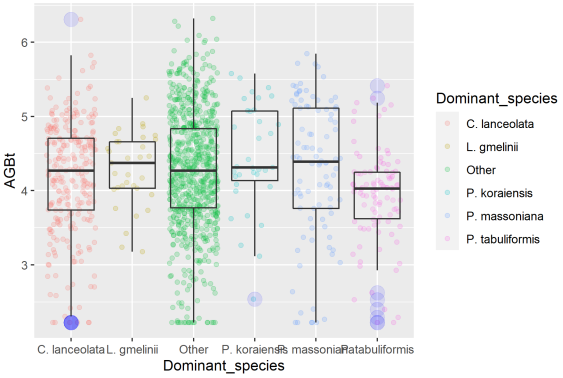


Figure 17. Box plot for Forest_type

Figure 18. Tukey HSD for Forest_type



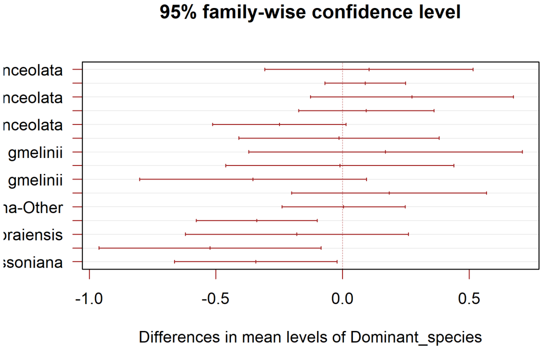
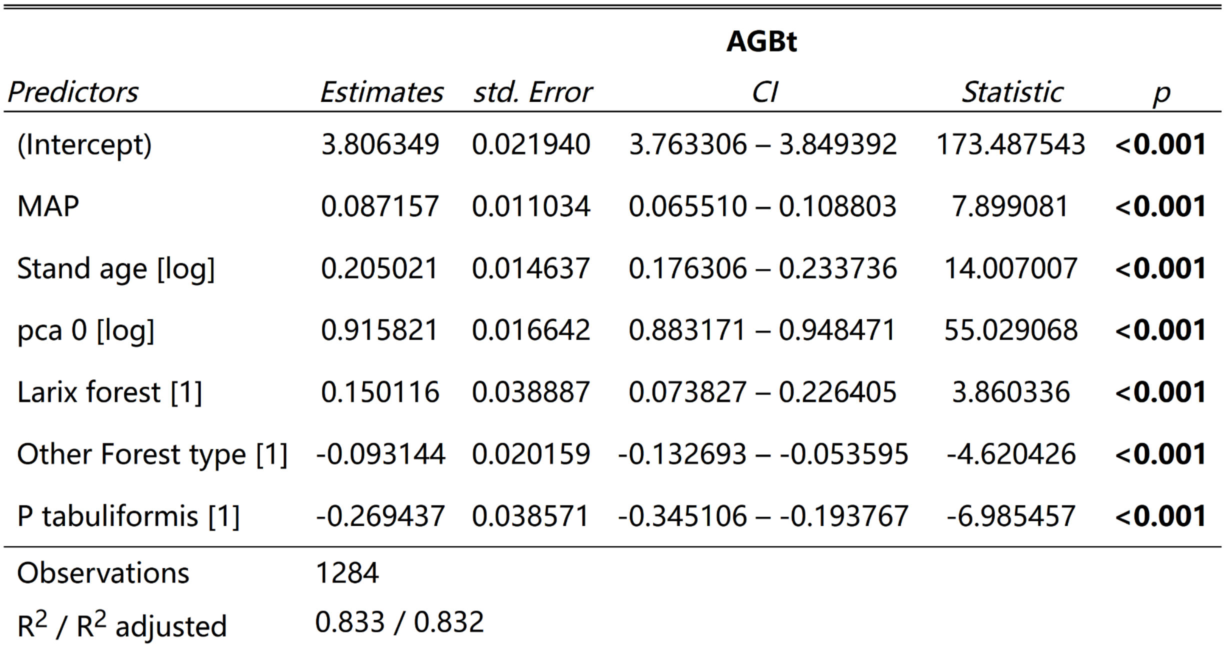


Figure 19. Box plot for Dominant_species

Figure 20. Tukey HSD for Dominant_species

The levels in Stand_origin appear to be significantly different based on the Tukey HSD plot, as all confidence intervals do not include zero. Their means also appear to be different from the boxplot. Soil_fertility shows an increase in mean as we increase the fertility level, but note that all the confidence intervals for adjacent levels contain 0 (e.g., for I-II, II-III, III-IV, and IV-V). Thus, we do not know where to separate the different levels. The Tukey HSD for Forest_type and Dominant_species is somewhat difficult to visualize, but the boxplots do display interesting information. The forest type with the highest mean appears to be 'Larix Forest' and 'Other', and all the means appear to be close. For the dominant species, P. tabuliformis shows a visibly lower mean than other classes. To prepare for model fitting, all four categorical variables were one-hot encoded. With all of the information above, we proceeded to fit multiple models and identified the best model.

4.8 Multiple Linear Regression



	AGBt				
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	3.806349	0.021940	3.763306 – 3.849392	173.487543	**<0.001**
MAP	0.087157	0.011034	0.065510 – 0.108803	7.899081	**<0.001**
Stand age [log]	0.205021	0.014637	0.176306 – 0.233736	14.007007	**<0.001**
pca 0 [log]	0.915821	0.016642	0.883171 – 0.948471	55.029068	**<0.001**
Larix forest [1]	0.150116	0.038887	0.073827 – 0.226405	3.860336	**<0.001**
Other Forest type [1]	-0.093144	0.020159	-0.132693 – -0.053595	-4.620426	**<0.001**
P tabuliformis [1]	-0.269437	0.038571	-0.345106 – -0.193767	-6.985457	**<0.001**
Observations	1284				
R^2 / R^2 adjusted	0.833 / 0.832				

Figure 21. Regression Output

We identified the best model as the one shown in Figure 21. All features are significant at α = 0.001, and the corresponding R^2 value is 0.832, adjusted. However, a pure R^2value is not representative of the performance of a model, as there is no threshold nor benchmark that defines a "good" R^2. A better way to determine a model's performance is to check for normality assumptions and perform validation.

4.9 Diagnostic Analysis

Linear regression can only be applied when the following three assumptions are met:

1. Normality: the errors are normally distributed
2. Independence: the errors are independent of one another
3. Homoscedasticity: the variance of the residual is the same for any values of the predictor

Q-Q plot and residual histograms were constructed to test normality.

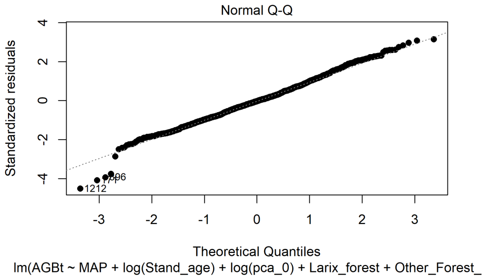


Figure 22. Q-Q Plot

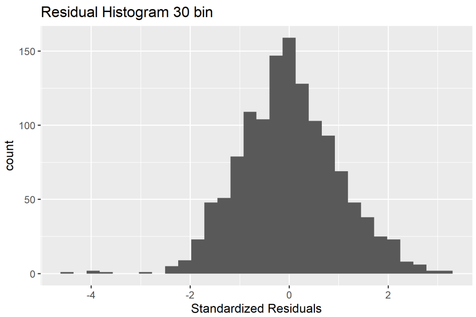


Figure 23. Standardized Residuals Histogram

As illustrated in Figure 22, the plot aligns quite well with the line representing the standard normal distribution. There appears to be a slight skew nearing the tails of the plot, potentially indicating that the data have more extreme values than would be expected if they truly came from a normal distribution. However, this skew is very minor as seen in the histogram in Figure 23. The 30-bin histogram confirms the normality of the residuals, with no significant points altering the approximately normal shape of the graph. The majority of the residuals fall into the bins around residual = 0 with a few outlying values on the negative side, corresponding to the outlying values on the Q-Q plot.

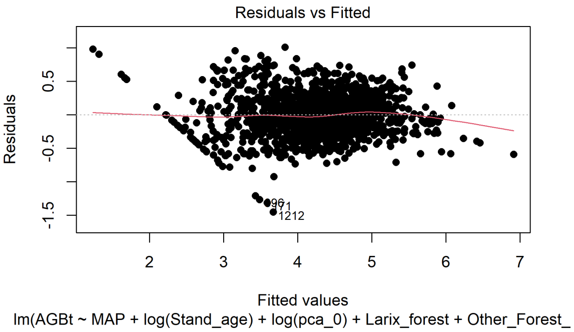


Figure 24. Residual Plot

The residual appears to be randomly and evenly distributed within the [-1, 1] range, with four outlying points on the lower end. They are centered around 0 and there is no fanning-out pattern that indicates heteroscedasticity, or other patterns that indicate non-constant variance. There is also no linear or nonlinear relationship present in the plot, meaning that the model should correctly encapsulate the relationship between the predictors and the response. One thing worth noting is that there appears to be a diagonal line forming on the left side of the residual plot with a few points. This behavior is a direct result of the winsorizing on outliers that we performed earlier. Since we floored the outliers at the whisker, it is expected that we get a result looking like the one shown above. This is by no means a violation of the randomness in the residual, because it accounts for such a small number of the total points (30 points, or 2.33% of the dataset, to be exact). Even without this knowledge about outlier handling, we wouldn't consider the residual plot in Figure 24 as a bad residual plot.

4.10 Prediction and Validation

To assess the prediction performance of the model, actual data from the validation set were plotted against predicted values from the model.

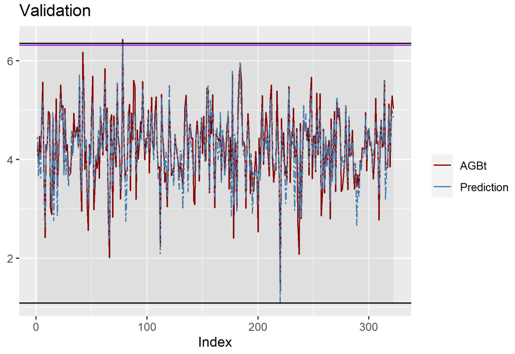


Figure 25. Validation Curve

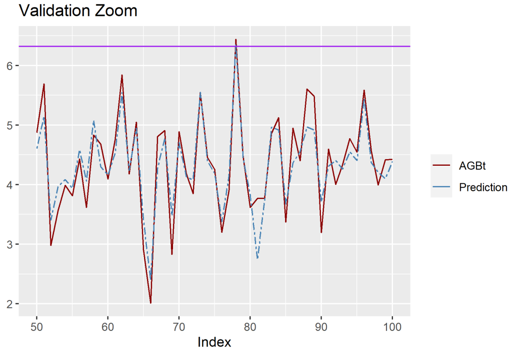


Figure 26. Validation Curve 6X Zoomed

Figure 25 reveals that the model predicts the trend of AGBt quite well, but sometimes the predicted results are more extreme than the actual data. The black horizontal band indicates the minimum to the maximum range of values predicted by the model, while the outlying spikes represent underestimated extrema. There appear to be only two such extrema over the entire prediction range. To improve visualization on the highly clustered curves in Figure 25, a 6X zoomed-in version of Figure 25 is constructed. As illustrated in Figure 26, the two curves follow a very similar trend with corresponding peaks and troughs in this portion of the validation curve.

5 Discussion

In our multiple linear regression model, the predictors are MAP, log(stand_age), log(pca_0), Larix forest, other forest type, and Pinus tabuliformis(name of a dominant species). The response variable is log(AGBt). For a researcher or government agency looking to assess the forest ecosystem, several implications arise from this study. MAP is the mean annual precipitation data. In an environment with more rainfall and snowfall, the ecosystem is in a good state for growth. The growth of the forest will result in the growth in the biomass. Therefore, it makes MAP a good valid predictor. Stand age is a description of forest lifespan. With an estimated value of 0.205, it can be interpreted as: with an ***n*** percent increase with stand age, the biomass will increase for (0.205*log(1+***n***)) units. This regressor is reasonable because stand age is also a great indicator of forest condition. The total biomass of a forest would only be high if the forest remains healthy for extended periods of time. The PCA principal component appears to be the most significant feature, which is to be expected as it is a linear combination of the biomass of the tree components. Even though PCA removes interpretability from the feature, we can regard this particular PCA component as a measure for overall tree health. Unsurprisingly, healthier trees produce a forest with higher biomass. Larix forest is one of the top 4 most common forest types in China. Including it in our model to predict the above-ground total tree biomass yields a good result. Pinus tabuliformis is a type of conifer that grows in northern China. Even in a cold and dry environment, Pinus tabuliformis is able to thrive. Since it favors high altitude and high latitude areas, this means that not many other trees and plants are able to survive the harsh environment where Pinus tabuliformis is in. Therefore, Pinus tabuliformis will exhibit a negative relationship with the overall AGBt. Reflecting upon the result of the model, the goal of prediction was successfully achieved.

Given the analysis above, there are still limitations that can be found in our dataset and analysis. At first glance at our dataset, it is obvious that the forest biomass dataset contains most of its study sites in eastern China, especially in southeastern China. There isn't much data collected for north-western China. Furthermore, we found that some data under the variables such as AGBf(Total aboveground living forest floor biomass) and BGBs(Total belowground shrub layer biomass) are missing, which makes them impossible for us to utilize. Granted more time and resources, the above limitations could possibly be resolved by collecting more data into the forest biomass dataset.

6 Appendix

A Data Availability

The original dataset can be found:
https://figshare.com/collections/Biomass_and_its_allocation_of_Chinese_forest_ecosystems/3306930

All code can be found in the GitHub repository listed below:
Enzoherewj/Forest-Biomass-Prediction (github.com)

B Contributions

All members contributed equally to this work. All members performed modeling, prediction, and wrote the paper.