Midterm

Ryan Hanson

11/2/2021

Abstract

Basal Area is used as a measurement of tree coverage. This analysis will use a dataset of cumulative basal area for lodgepole pines in square feet per acre as a response variable for various explanatory variables to create a predictive model and identify important environmental factors. I was able to create a random forest model with predictive power that can identify that the degrees from north for the slope is the largest environemntal factor.

Introduction

Lodgepole pine trees can have a wide variation of basal area, depending on numerous environmental factors. The goal of this analysis is to create a model to predict the cumulative basal area of lodgepole pines in square feet per acre and analyze which explanatory variables have the largest predictive importance. We have a dataset containing variables longitude, latitude, slope, aspect, and cumulative basal area of lodgepole pines. More detailed descriptions are listed below in table 1. Our dataset has 114 observations containing these parameters, and an additional 78 observations containing everything except for the lodgepole's cumulative basal area.

Since gathering data for our basal area response variable can be time consuming and impractical, this model will allow us to estimate, within reason, the cumulative basal area of an acre given our other factors. We will use the first 114 rows to create and cross validate our model, then use the model to predict the lodgepole basal area for the last 78 observations.

Table 1: Descriptions of the variables of most importance

Parameter	Description
LON	Longitude
LAT	Latitiude
Slope	Average slope of plot in degrees
Aspect	Counterclockwise degrees from north facing for slope
ELEV	Elevation of plot centroid in feet
Lodgepole	Cum basal area of lodgepole pines in ft^2/acre

Exploratory Data Analysis

In order to determine which model will be most effective, we must understand our data. Figure 1 contains a plot of our cumulative basal area displayed by latitude and longitude. Each dot on this graphic represents an acre measured, and the color represents which cumulative basal interval the acre falls under. One concern with our data is the need to account for spatial autocorrelation, since our terms need to be decorrelated. Therefore, our model must take that into consideration in order to be reliable.

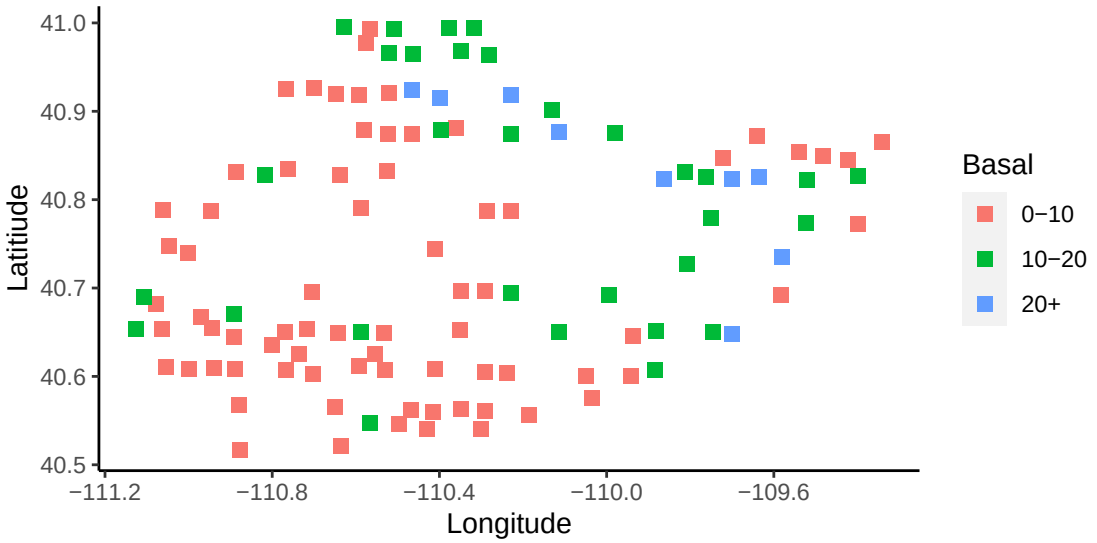

Figure 1: Location of acres, colored by cumulative basal area

Next, we compare scatterplots containing each explanatory variable compared to our lodgepole response variable. One of the key things to note from the scatterplots is that there is not a linear relationship between any of the explanatory variables and the response variable. This is shown in figure 2 where I included lodgepole area compared to aspect, slope, and elevation. Because the data lacks linearity, we will need to choose a model that can appropriately handle that, or else we will build an extremely unreliable model. This likely means we will need a non-parametric method.

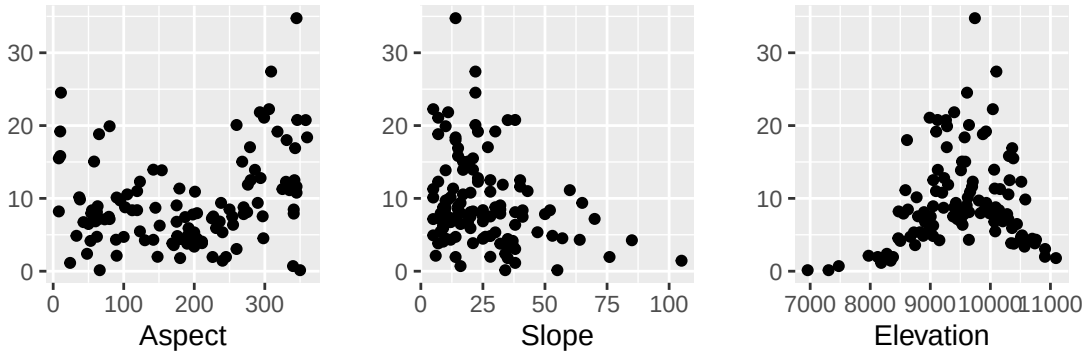

Figure 2: Scatterplots of aspect, slope, and elevation compared to lodgepole basal area

Models Selection

I looked at multiple different models in order to determine which would be the most effective in completing the goals of our analysis. In order to verify our initial EDA, I ran a linear model on our data. This resulted in an extremely low R^2 of .25. This model also proved that it didn't meet our assumptions of equal variance and normally distributed residuals, so we can conclusively reject the use of this model without some potentially extreme transformations.

The next model I created was a tree model. A tree model is a non-parametric model and therefore far more adaptable to data that is non-linear. It allows for some interpretation, which is useful, but will have a

weaker predictive power compared to some of my other models. Once it builds out the model, I can prune it back to avoid overfitting the data. Figure 3 shows what a pruned tree looks like, and when each condition is followed down the branches, I end up with an approximate expected value for one specific area given its environmental and spatial conditions.

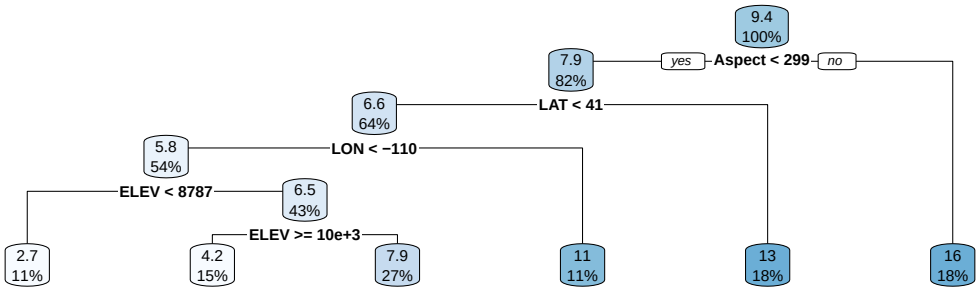

Figure 3: Tree model

Finally, I created a random forest model. A random forest model is a complex tree structure, where I can specify the number of trees (B) I pass our point of interest (x_0) through and then compute the mean of the expected value ($\hat{y}(x_0)$) from each individual tree (b, see figure 3 for example of one tree). I use that number as my prediction for the acre given its environmental and spatial conditions. The formula for this computation is shown below in formula 1. The strengths of a random forest include a high accuracy in prediction and an innate ability to handle spatial autocorrelation since it's a non-parametric model. It also removes any interaction between variables as each tree randomly selects four of the five variables and and calculates a value using those four parameters. This allows me to justifiably leave all parameters in our model without any complications. Random forests do not have any assumptions about the data that need to be met in order to be utilized.

$$\hat{y}(x_0) = 1/B \sum_{b=1}^{B} \hat{y}^b(x_0)$$

x_0 Represents our point of interest

$\hat{y}(x_0)$ Represents the predicted value for our point of interest

B Represents the total number of trees in our random forest

b Represents the bth tree which our point passes through

$\hat{y}^b(X_0)$ Represents the prediction of y for the bth tree

\ The histogram in figure 4 shows that the data is unbiased and the residuals are normally distributed. The primary way to measure the effectiveness of a model is root mean square error (RMSE). The RMSE is a measure of the standard deviation of the residuals. In order to caculate RMSE, I split the data into a train and test set. I built all my models on the train set, then ran the test set and train set through our model, calculating the RMSE and in-sample RMSE. The RMSE in our forest model was 3.75 and the in-sample RMSE was 4.39. The RMSE for our tree was 5.21 and the in-sample RMSE was 4.64. Since the linear model doesn't meet the assumptions, the RMSE isn't useful for determining predictive capability. The random forest clearly outperforms our other models.

Another important measure of my model is bias. Bias is the average difference between our predicted and observed values. The bias for my tree model was -1.14 and my random forest was .96. The bias shows that both of the models are close enough to 0 to be justified as unbiased.

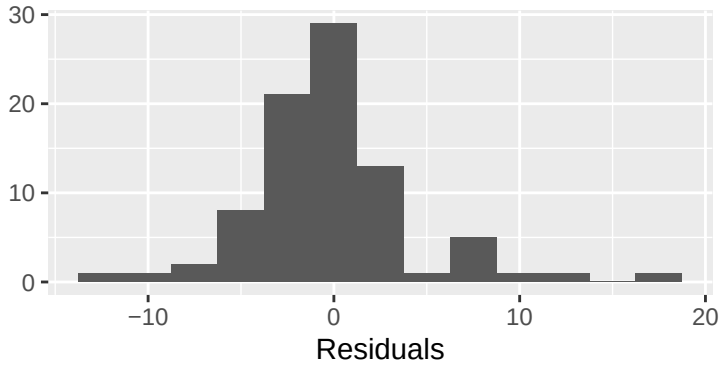

Figure 4: Histogram of the Residuals

Results

The first question I set out to answer was which variables have the largest effect on cumulative lodgepole basal area per acre. With a random forest model, I am able to rank variables by order of importance. Figure 5 below shows that aspect is the most significant variable, with latitude, longitude, elevation, and slope next in importance. The downside of a random forest analysis is a lack of quantifiable interpretation. So although I can order each variable's importance and see their rank relative to other variables, I can't give a specific numeric effect on the response variable. I can only conclude that the variable aspect has the largest effect in determining cumulative basal area for lodgepoles.

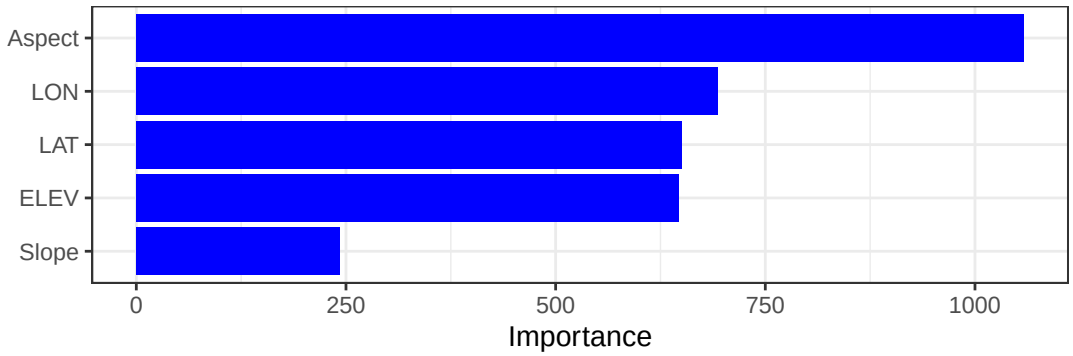

Figure 5: Importance of each variable

The second question involved prediction of basal area for the unmeasured acres. The purpose of this model is to accurately use the given features to predict basal area without the tedious work of physical measurement. I used the random forest model to predict lodgepole basal area in our unmeasured areas. Figure 5 shows an overlap of the initial data spatially, represented by the hollow squares, and the new predicted values represented by the solid squares. We are able to estimate uncertainty on these predictions as well, but they will not be able to be represented in a visual representation.

Conclusion

My model proved to be effective in meeting the goals of the analysis. I was able to rank importance of the explanatory variables and create a model with predictive power. This model will eliminate the time needed to physically measure each tree, and allow prediction using easy to measure characteristics. The downside of

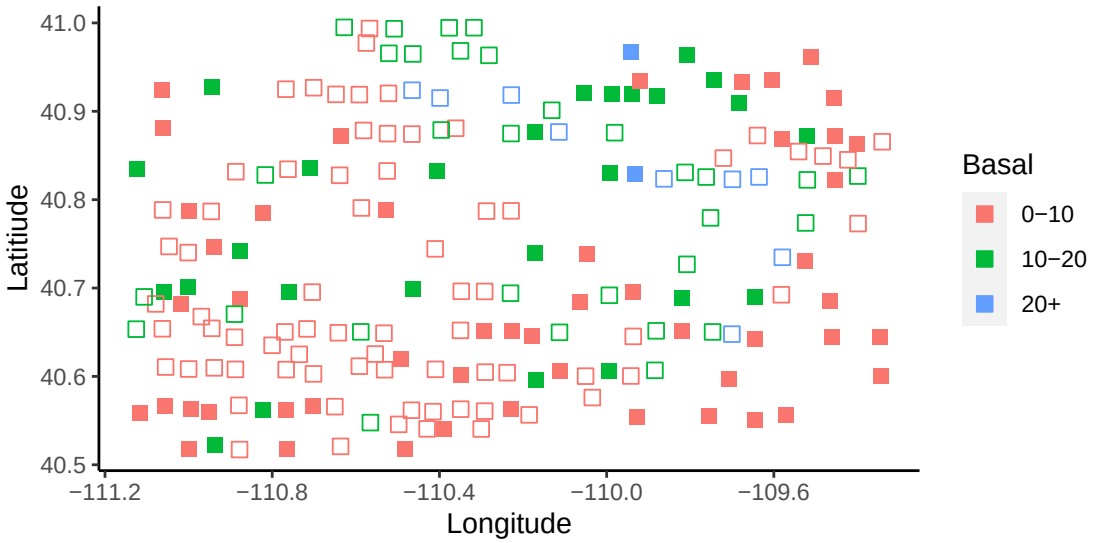

Figure 6: Basal area measurements in dataset (hollow squares) and unmeasured values predicted by model (solid squares) graphed spatially

the model is that although I can rank variable importance, I can't give an interpretable value to measure the rate at which the variables affect the response variable. Another downside is that this model isn't extremely adaptable to other locations, since the latitude and longitude is relative to this area. An interesting next step could involve finding multiple forests with differing climates and creating a more robust model that's not dependent on spatial data, but wholly dependent on environmental factors.