Camilla Handley

STAT 536 Midterm

Lodgepole Basal Area Analysis

**Introduction**

In the Uinta National Forest there is an infestation of pine beetles. The pine beetles feed and tunnel into pine trees and eventually kill the tree. In lieu of this infestation, the Forest Inventory Analysis (FIA) wanted to investigate what kind of environments are conducive to lodgepole pine growth, or in other words, what factors other than the beetles affect Lodgepole pine growth. The data that we will use for this analysis was collected by the FIA team at various locations in the Uinta National Forest. For each Lodgepole pine, the longitude and latitude coordinates were recorded as well as the average slope of the plot of land where the tree is located, the aspect (counterclockwise degrees from north facing), elevation (in feet), and cumulative basal area of the tree. We will use this data to see if any of these factors have significant effects on Lodgepole pine growth based on its basal area. In addition, there were some locations where the FIA could not go and record the basal area. Since the coordinates, slope, aspect and elevation are known for these locations, we want to accurately predict the basal area at these locations.

**Exploratory Data Analysis**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedGiven that this data was collected from one general area, we have reason to believe that the observations might be spatially correlated. This means that observations that located close together are correlated. We can see evidence of this in Figure 1, as some darker points are clustered together. We can also check if there is spatial correlation or not by looking at a variogram (Figure 2). If there were no spatial correlation, the points would form a relatively straight line indicating that as observations get farther away, the semi-variance stays constant. However, the points have a general upward trend, indicating spatial correlation in the data.

***Figure 1: Basal Area by Location Figure 2: Variogram***

Having correlated data would violate our independence assumption in linear regression. If we simply ignored the correlation and performed the analysis with multiple linear regression, our standard errors would shrink, causing our predictions and inference to be inaccurate. In addition to correlation in the data, there are also signs of non-linearity. Figure 3 shows Added Variable plots from a simple linear model with log(Basal Area) as the response variable. There definitely some curvature in Elevation and Aspect. Given that these variables do not have a linear relationship with Basal Area, we cannot use a simple linear model to analyze this data. Our results would be incorrect because the model would not fit the data. We will discuss how to deal Chart, diagram, scatter chart

Description automatically generatedwith these problems later in the analysis.

***Figure 3: Added-Variable Plots***

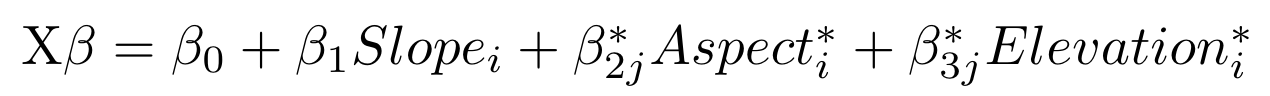
**Statistical Model**

To find an adequate model for this data, we need to account for the non-linearity and the correlation. There were also some signs of non-equal variance, so we decided to base our models off of log(Basal Area) as the response. This transformation helped normalize the variance of the residuals. For non-linearity, various methods were tried, including transformations, natural splines, polynomials and GAMs (generalized additive models). In terms of AIC, the GAM performed the best, however, there is no way to implement correlation in a GAM, so we decided on natural splines, which was the second best performing model. We chose AIC as our criteria because of the predictive nature of our analysis goals. Natural splines split the data into different sections and fit twice continuously differentiable cubic polynomials. This just means that it will fit a smooth line across the data, without any breaks or sharp turns. Each section is linear, therefore fixing the non-linearity problem. However, natural splines should not be used to predict outside of the range of the data because they are linear outside of the data and do not reflect the true behavior of the data. However, the data that we have been asked to use to predict is in or barely outside of the range of our observed data, so there shouldn’t be a problem. By way of cross validation, it was found that a natural spline with 2 degrees of freedom performed best in terms of Root Mean Square Error (RMSE) for both Aspect and Elevation.

To deal with the correlation, a few different correlation structures were implemented using the gls function in R. When compared to Spherical correlation and Gaussian correlation, Exponential correlation performed the best in terms of AIC. In addition, since there is the possibility of two observations recorded at the same or very similar locations (coordinates) we added a nugget to our model. This led us to form our final model, which is defined below.

Text

Description automatically generatedDiagram, text

Description automatically generated with medium confidenceIn this model, **Y** is the vector of log(Basal Area). Slopei represents the Slope of the ith observation. is the intercept term, or the expected log(Basal Area) of a tree with slope, aspect and elevation equal to zero. represents the average expected increase in log(Basal Area) as the slope increases by one, holding everything else constant. What has been identified as. actually a collection of two coefficients. The value of this term will change depending on the value of Aspect. Aspect\* is a cubic recentering of the Aspect variable which is the application of natural splines in our model. The interpretation can be applied to but for Elevation instead of Aspect. This model assumes correlated errors, where the correlation is captured and defined by **R**. is defined as the “range” parameter – as increases (at a fixed distance) correlation increases as well. is the Euclidean distance between two points. captures the nugget effect – so that when the Euclidean distance is equal to zero (i.e. two observations have the same coordinates), it gives added variance and allows sampling variability. Nuggets also help stabilize estimation. We use this model under the assumptions of independence of the data (after decorrelating the residuals), linearity, normality of decorrelated (or standardized) residuals and equal variance of decorrelated residuals.

**Model Justification**

Graphical user interface, chart, application

Description automatically generated In order to continue to use this model for inference and prediction, we need to show that the assumptions (mentioned previously) are justified. Figure 4 shows added variable plots from a linear model after the natural splines and log transformation were applied. With no obvious curvature on any of the plots, the linearity assumption is justified.

***Figure 4: Added-Variable Plots from Transformed Model***

Figure 5 shows the variogram of the standardized residuals from our final model. We can see that there is not an obvious upward trend in the dots but rather a relatively flat line. This shows that our model has adequately accounted for the correlation, as the semi-variance does not increase as distance increases. Thus we can continue under the assumption of independence in the data. Figure 6 shows a histogram of the standardized residuals from our final model. We can see that the residuals have an approximate normal distribution centered around zero and we can Chart, scatter chart

Description automatically generatedcontinue under the assumption of normality.

Chart, histogram

Description automatically generated

***Figure 5: Variogram from Final Model Figure 6: Histogram of Standardized Residuals***

Chart, scatter chart

Description automatically generatedLastly, to justify equal variance, we will look a plot of the fitted values (untransformed) and the standardized residuals. There is no pronounced clustering, and the residuals seem to be evenly distributed, except for a few outliers from high values of basal area. We will continue under the assumption of equal variance.

***Figure 7: Fitted Values vs. Residuals Plot***

**Model Evaluation**

To evaluate the predictive capability of our model, we performed a leave-one-out cross validation study. The results are listed in the table below.

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Bias | -0.0458 |
| Coverage | 0.9825 |
| RMSE | 1.409 |

Given that the Lodgepole data has a standard deviation of 6.239 and a range of 34, we have a very good RMSE score. We are predicting well within one standard deviation. Our bias is very small and negative, indicating that on average we are slightly underpredicting. Our model has a coverage of 98.3%, indicating that 98% of our prediction intervals contain the observed basal area. Overall, our model has very good predictive capability. To assess model fit, we calculated a “Pseudo” R-squared. This is calculated by squaring the correlation between the observed values and the fitted values. We have “Pseudo” R-squared of 0.6376, indicating adequate model fit.

**Results**

One of our research goals was to find what kind of environments are conducive to Lodgepole pine growth. To answer this question, we will look at the coefficients from the model, displayed in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Covariate** | **2.5%** | **97.5%** | **Estimate** | **p-value** |
| Slope | -0.0088 | -0.0043 | -0.0065 | < .0001 |
| Aspect - 1 | -0.6794 | -0.2624 | -0.4798 | < .0001 |
| Aspect - 2 | 0.7775 | 0.9959 | 0.8867 | < .0001 |
| Elevation - 1 | 8.0555 | 8.9035 | 8.4795 | < .0001 |
| Elevation - 2 | -0.0571 | 0.3203 | 0.1316 | 0.1745 |

Given these results, we can conclude with 95% confidence that as the slope of the plot increases by one degree, we would expect an average decrease of between -0.008 and -0.004 in log(Basal Area), holding all else constant. Overall, slope has a negative effect on Basal Area, although it is very slight (i.e. as slope increases, basal area decreases slightly). Rather than trying to interpret the coefficients for Elevation and Aspect, we will use our model to visualize the effect of these covariates. Figure 8 shows a graph of predicted values (and their prediction intervals) for a general range of elevation values, while holding everything else constant at their average value. The blue bands represent the prediction intervals while the grey points show the observed data. We can see that as the elevation has both a positive and negative effect. Up until about 9500 feet, as elevation increases, so does basal area. However, after about 9500 feet, basal area starts to decrease as elevation increases. The optimal elevation for maximum basal area would be between 9000-10000 feet.

Chart

Description automatically generated

***Figure 8: Elevation Predictions***

***Chart

Description automatically generated***Similarly, Figure 9 shows a graph of predicted values and intervals for all possible values of Aspect (0 – 360), while holding everything else constant at their average value. The blue bands represent the predictions and the grey dots are the observed values. We can see that plots that are closer to being north facing (i.e. Aspect closer to zero or closer to 360) have higher predicted values of basal area and south facing plots (Aspect 180) have lower predicted values of basal area. We can conclude that plots of land that are north facing (or close to it) are optimal for Lodgepole pine growth.

***Figure 9: Aspect Predictions***

Chart, scatter chart

Description automatically generatedChart, line chart, histogram

Description automatically generatedWe used this model to predict basal area for the plots where the FIA could not go. Figures 10 and 11 provide a visual representation of the results. Figure 10 shows a map based on the coordinates where the color represents the predicted basal area (with darker blue points corresponding to smaller basal area). Figure 11 is a graph of the predictions and their intervals with Elevation along the x-axis. We have also attached a data set with the predictions.

***Figure 10: Predictions by Coordinates Figure 11: Elevation and Predictions***

**Conclusion**

We were able to build a model that fit the data fairly well and had great predictive capability. This model dealt with the non-linearity of some of the covariates as well as accounted for the spatial correlation between observations. We were able to confidently predict basal area for the trees at the locations where the FIA was not able to go. We concluded that Lodgepole Pine growth is affected mainly by the elevation and aspect of the plot where the tree is located. Based on our results, a plot of land that would be considered very conducive to basal area growth would be north facing, have an elevation of between 9000 and 10000 feet and a slope of zero. If we needed to plant more Lodgepole pines, it would be best to plant them in these sorts of environments. We acknowledge that these may not be the only factors that affect Lodgepole Pine growth. If another opportunity presented itself to collect data, it might be helpful to know where populations of pine beetles are located, as pine beetles are detrimental to Lodgepole Pine health. It would also be beneficial to collect more data from more trees, as having more data usually results in a more accurate model. We would consider information about what other kinds of plants are growing near each Lodgepole pine, to see if certain species are detrimental or beneficial to growth. We know that this may not be the best fit model possible. Various techniques were tried and compared in the search for this model, but not all possible combinations nor techniques were tried. Exponential correlation is a special case of a broader class of correlation function called the Matern correlation. It could be beneficial to try other correlation functions in this model.