Stat 536: Midterm

Madeline Morris

Brigham Young University

Abstract. This analysis uses non-linear and spatial correlation methods to model the affect of environmental characteristics on lodgepole pine basal area measurements. The purpose of the study was to understand what type of environment is best for lodgepole pine tree growth as measured by basal area and predict for trees that weren't able to be measured. From this analysis, I concluded that the best environment for lodgepole pine trees to grow is an area with low slope and high elevation and high aspect. Additionally, I can use this model to predict for trees that are not able to be measured. Through testing, these predictions seem fairly accurate. To increase the R^2 value, I would recommend adding additional covariates that could explain more about what affects lodgepole basal area. These covariates could include information about the type of soil or amount of precipitation that area receives, although this information may be difficult to collect.

1 Introduction and Data Summary

Uinta National Forest is a national forest located in north central Utah. It is a gorgeous forest with a variety of pine trees. This forest is a peaceful escape for travelers to enjoy hiking, camping, mountain biking, snowshoeing, and more. Part of what has kept this forest healthy and thriving is the mountain pine beetle. These pine beetles are insects which burrow and reside in pine forests, primarily in the western United States. These pine beetles are beneficial to forests because they speed up the growth of younger, healthier forests by attacking weaker trees. However, in recent years the western United States has experienced unusually warm and dry summers. This has led to a dramatic increase in pine beetle numbers which then led to a greater decline in tree populations. Due to this, environmentalists seek to understand what factors promote the growth of trees, so they know where to plant new trees.

The Forest Inventory and Analysis (FIA) Program of the U.S. Forest Service performs studies to evaluate the condition of these forests. This service works to understand how forests will appear in the future, to understand if current forest management will lead to a sustainable forest for the long term. The data collected includes, the longitude coordinate of the plot, the latitude coordinate of the plot, average slope of the plot in degrees, the aspect or counterclockwise degrees from north facing, the elevation of plot centroid in feet, and the cumulative basal area of each tree in square feet/acre. Although FIA provides beneficial data, it isn't always complete. To study the condition of trees, someone has to physically go to the forest and measure each tree. Thus, the cumulative basal area of some trees remain unknown due to difficulty in traveling to the tree.

This study will focus solely on lodgepole pine trees in the Uinta Forest. Using data from FIA, I will seek to understand how a lodgepole pine basal area is affected by its environment. Essentially, I hope to understand which environmental factors promote healthy lodgepole pine growth. Additionally, I will use the data given to create a model with which I can then predict the basal area for trees where FIA wasn't able to go.

Figure 1 gives an initial look to the data. As can be seen by the numerous grey dots, there are many missing observations in this dataset. Removing the missing values provides a clearer view as to how the lodgepole basal area changes by location.

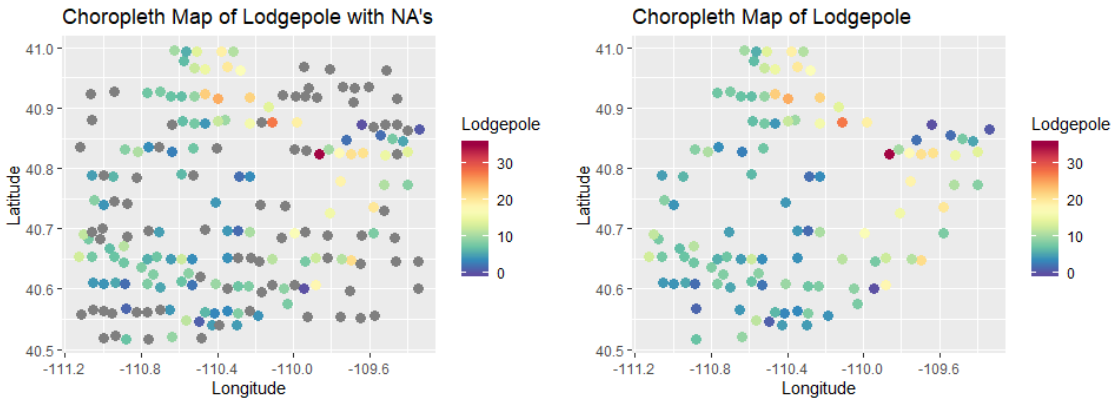

Fig. 1. Choropleth Map of lodgepole basal area by location. This clearly shows that there seems to be spatial correlation as there are clusters of similar values across the plot.

From the plots in Figure 1, it seems clear that there is spatial correlation, meaning that trees with larger basal areas are clustered together and trees with smaller basal areas are clustered together. The variogram in Figure 2 shows that the variance of the difference of the response variable between one location and a second location changes with distance. If there was no spatial correlation, these points would have a flat trend.

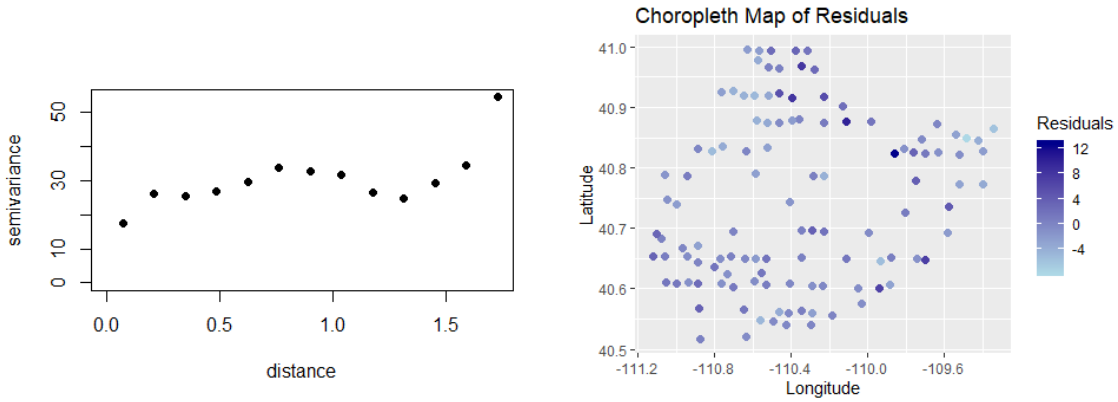

Fig. 2. Variogram and Choropleth Map of residuals from a linear model. Both of these plots indicate that their is spatial correlation.

I also fit a simple linear model and plotted the residuals by location as shown in Figure 2. There are clusters of dark and light points, also indicating that the residuals are spatially correlated. The consequence of ignoring this spatial correlation would cost me predictive ability. The predictions would be accurate, but the prediction intervals would be too wide. Accounting for the correlation in the data will allow the predictions to be much more accurate.

I was also interested in exploring the relationship between all the other covariates and lodgepole basal area. Table 1 below shows the correlation between each covariate and lodgepole basal area. None of these values are close enough to 1 to indicate a strong linear relationship between the covariate and lodgepole basal area. Figure 3 shows the original data for lodgepole basal area and slope, aspect, and elevation.

Latitude	Longitude	Slope	Aspect	Elevation
0.3	0.35	-0.27	0.24	0.14

Table 1. Correlation values between lodgepole basal area and each covariate in the dataset. None of these values indicate a strong linear relationship.

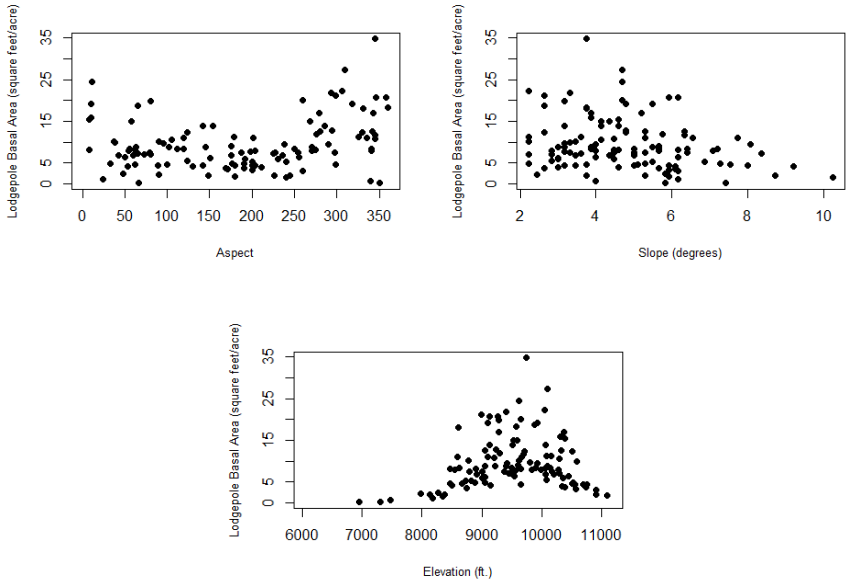

Fig. 3. Exploration of original data plotted against lodgepole basal area.

The plots in Figure 3 show that aspect and elevation have non-linear relationships with lodgepole basal area. Longitude and latitude had a linear relationship with lodgepole basal area, so I have chosen not to include their plots. Although slope doesn't seem to have a very strong linear relationship with logdepole basal area, I decided after testing that this relationship is linear enough to include slope without any transformation. If we included the non-linear covariates in the model without accounting for non-linearity, then any predictions or confidence intervals may be biased and inaccurate. Essentially, if the data is curved and a linear line is fit, then the model won't match the data and all analysis after fitting the model will not be valid. I will therefore use non-linear methods to fit a linear model with variables to account for this issue.

Accounting for non-linearity can lead to further complications. By accounting for non-linearity through the use of splines or other known techniques, I might be introducing collinearity to the model or I may be overfitting. For instance, if too high of an order is fit for a spline, the model may be overfitting the data, causing inaccurate and/or biased predictions. Additionally, if I introduce too much collinearity into the data then any predictions would have the same accuracy, but the prediction intervals would not be valid because standard error estimates would be incorrect. Thus, I must be careful when incorporating a non-linear effect into the model, but also acknowledge that this will have an effect on the standard errors and predictions.

Thus, to account for the spatial correlation and non-linearity in the data, I will use non-linear methods to fit a linear model with variables to account for non-linearity in the data, as well as incorporate a correlation structure into the model to account for the correlation between location.

2 Statistical Models

After much testing, I concluded that to produce reliable results that accounted for the non-linearity and spatial correlation in the data, I needed to transform the response variable. I decided to do a square root transform on lodgepole basal area. Thus, moving forward, I will be creating a model that models the square root of lodgepole basal area. However, all predictions will be on the original scale. For reference, aspect, slope, and elevation plotted against the transformed response variable is shown in Figure 4

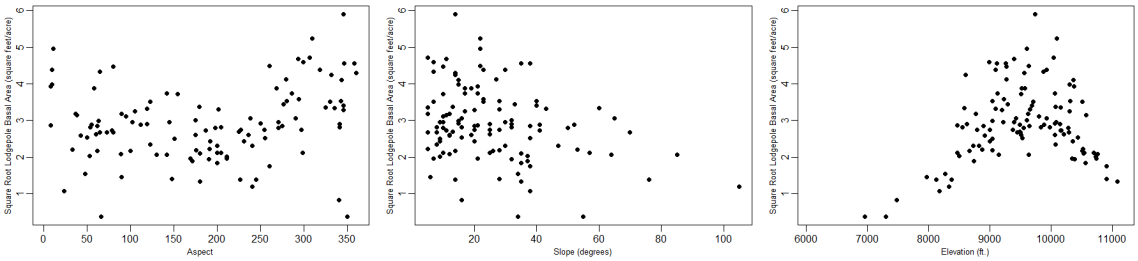

Fig. 4. Exploration of covariates plotted against transformed lodgepole basal area.

To account for the non-linearity in the data, I re-defined the variables aspect, and elevation to conform to a linear model. The variable aspect is a circular variable. Meaning, that aspect is a measurement of how many degrees from north facing (0-360 degrees). Thus, we would expect a tree with an aspect measurement at 350 degrees would have a similar lodgepole basal area as a tree with an aspect measurement of 5 degrees. After trying a few transformations, I decided to use a polynomial basis expansion. Polynomial regression is simply a linear regression with a basis function expansion on a nonlinear variable. Using a polynomial regression, the nonlinear term is expanded by degrees. These expanded terms are then orthogonalized to ensure that there is no collinearity between the expanded terms. Essentially, this means that a regression model is fit on each expanded term with all the previous expanded terms used as explanatory variables to extract the unique contribution of each subsequent polynomial term. This model raises some variables to high powers, thus creating very large numbers. Thus, it is essential to center and scale the expanded terms to allow for easier computation.

Polynomial regression is advantageous because it can fit a wide range of curvature. It also allows the use of all the same linear model methods for estimating model fit and predictive ability. One downside to polynomial regression is that it is highly sensitive to outliers. In addition there are fewer model validation tools to detect outliers in nonlinear regression than there are for linear regression. Another disadvantage is that polynomial regression often fits terribly at the edges of the data. Thus extrapolation is even more dangerous with polynomial regression than with linear regression. However, in this analysis, all predictions will be roughly within the range of the original data, so extrapolation isn't a severe issue. Polynomial regression relies on the regular linear regression assumptions. Additionally, to use polynomial regression, you must specify the degree to raise the x variable to. If too high of an order is fit, the model may be overfitting the data, causing inaccurate and/or biased predictions.

To decide the number of expanded terms of highest degree of expansion, I used cross validation to identify the model that minimized BIC. Figure 5 below shows the BIC for increasing expansion terms. The BIC was minimized at 2 degrees of polynomial order. Thus, I decided to use a 2nd order degree polynomial expansion on aspect for this model.

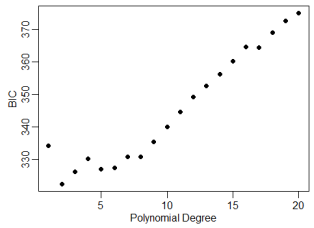

Fig. 5. BIC for each degree when fitting polynomial expansion on aspect.

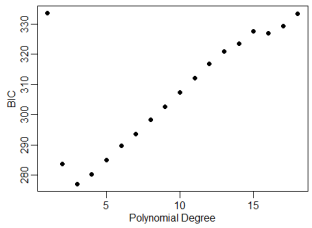

Fig. 6. BIC for each degree when fitting polynomial expansion on elevation.

To adjust for the non-linearity in elevation, I again chose to use polynomial regression. To decide the number of expanded terms of highest degree of expansion, I used cross validation to identify the model that minimized BIC and AIC. Figure 6 shows the BIC for increasing expansion terms. The graph for minimizing AIC looked almost the same. The BIC and AIC was minimized at 3 degrees of polynomial order. Thus, I decided to use a 3rd order degree polynomial expansion on elevation for this model.

After fitting a 2nd order degree polynomial to aspect and a 3rd order degree polynomial to elevation, I used variable selection to select which variables to keep in my model. Both AIC and BIC produced the same results. Additionally, I used the exhaustive subset method to look at every possible combination of explanatory variables. The best model included all the variables, so I decided to keep all the variables, (longitude, latitude, slope, aspect, and elevation) in the model.

Additionally, I tested several different correlation structures to account for the correlation in the data. After testing exponential, Gaussian, and spherical correlation structures, I found that the spherical structure resulted in the smallest AIC and BIC values. Therefore, I decided to use the spherical structure in my model to account for the correlation. Further, I decided to use a nugget in my correlation structure because this also resulted in the smallest AIC and BIC values.

Spherical correlation structures are easy to compute because there is only one parameter that must be estimated from the data. In my model, there are two parameters to be estimated since I included a nugget. This makes computation and interpretation simple. One disadvantage to this correlation structure is that it only builds in correlation if the sample points are within a certain distance from each other. It is similar to a moving average correlation in that way. Thus, if this correlation isn't matched in the data, I may be losing some information that would be maintained with a different correlation structure. However, this model and correlation structure minimized AIC and BIC, so I feel confident proceeding with this model as shown below.

$$sqrt(y_i) = \beta_0 + \beta_1 Longitude_i + \beta_2 Latitude_i + \beta_3 Slope_i + \beta_4 Aspect_i + \beta_5 Aspect_i^2 + \beta_6 Elevation_i + \beta_7 Elevation_i^2 + \beta_8 Elevation_i^3 + \epsilon_i, \qquad \epsilon \sim N(0, \sigma^2((1-\omega)R + \omega I)). \tag{1}$$

$sqrt(y_i)$ = the square root lodgepole basal area in square feet/acre for the ith tree.
β_0 = the average square root lodgepole basal area. Essentially, this is the expected square root lodgepole basal area when not accounting for the change in slope, aspect, or elevation and not accounting for any correlation in the data.
β_1 = the amount that square root logepole basal area will increase for a one unit increase in longitude.
β_2 = the amount that square root logepole basal area will increase for a one unit increase in latitude.
β_3 = the amount that square root logepole basal area will increase for a one unit increase in slope.
$\beta_4 - \beta_8$ = the effect that aspect and elevation have on square root lodgepole basal area. Each value of aspect or elevation is raised to the specified power and then orthogonalized as described earlier. One unit increases in those variables (with all else held constant) cause an average increase in square root lodgepole basal area equal to the respective β value.
ϵ_i = The distance from the observed lodgepole basal area measurement and the predicted lodgepole basal area measurement for each tree.

σ = The standard deviation of the errors.
ω = The nugget in the model. This allows for added variance when $|d_1 - d_2| = 0$. This allows for samples at the same exact location to not be perfectly correlated with each other. This is estimated from the model.

The R matrix is an 114×114 covariance matrix with 1's along the diagonal and the correlations between two observations at different locations along the off diagonal. These correlations are defined as $\rho(s_i, s_j)$, meaning the correlation between location i and j. For this model, I used a spherical correlation function. This correlation function is defined as:

$$\rho(s_i, s_j) = \begin{cases} 1 - 1.5\left(\frac{\|s_i - s_j\|}{\phi}\right) + 0.5\left(\frac{\|s_i - s_j\|}{\phi}\right)^2 & \text{if } \|s_i - s_j\| < \phi \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where ϕ is the range parameter which is estimated from the model. Spherical correlation structures account for correlation only if the data points are so far apart. After a certain distance, samples are not correlated at all. This is called a taper correlation which is similar to a moving average correlation.

3 Model Assumptions

To use this model, several assumptions must be met. First, that the relationship between the response and explanatory variables is linear. This is obviously not met, but the use of polynomial basis expansions accounts for this. Second, I am not assuming independence in the residuals but I am incorporating this into the model by using a spherical correlation matrix for the errors to account for the dependence. Third, I assume that the standardized and de-correlated residuals of the response are normally distributed. And fourth, I assume there is equal variance in the standardized and de-correlated residuals. If linearity is not met or accounted for, then any predictions or confidence intervals may be biased. If the residuals are not normally distributed then confidence and prediction interval calculations would be invalid since they are based on the assumption of normality. If equal variance isn't met, the standard errors would be incorrect. If independence is ignored, the standard errors will be too small and any prediction or confidence intervals will not be accurate. Thus, it is important to verify each of these assumptions for both models.

To continue with this model, linearity must be met. However, it was clearly shown that aspect and elevation do not have a linear relationship with lodgepole basal area. This non-linearity was accounted for with the use of polynomial basis expansions. Figure 7 shows how the polynomial basis expansions fit to the data. These plots indicate that these polynomial expansions fit the data well. Reiterating the fact that the variable, aspect, is circular, I would expect the lodgepole basal area to be similar for a tree with aspect values near 0 and near 360. The polynomial expansions seems to model this trend as the curve goes up at both endpoints of the data, thus I feel that the polynomial expansion fits the data well.

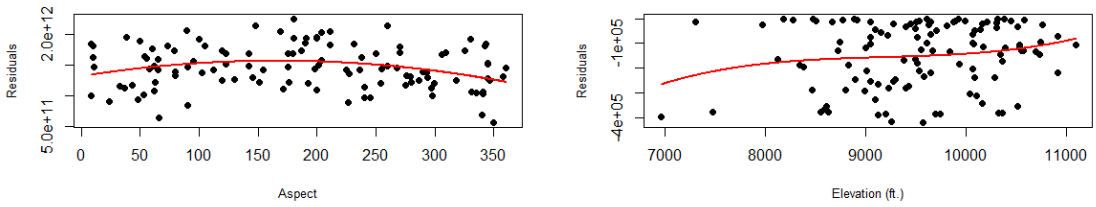

Fig. 7. Fitted model plotted against aspect or elevation and the residuals from the data after removing the affect of all other variables. These plots show that the polynomial expansions fit each variable very well.

To show that the linearity assumption is met, we can look at added variable plots of each covariate as shown in Figure 8. It appears that there is no non-linear trend in any of these plots, thus we can assume that linearity is met.

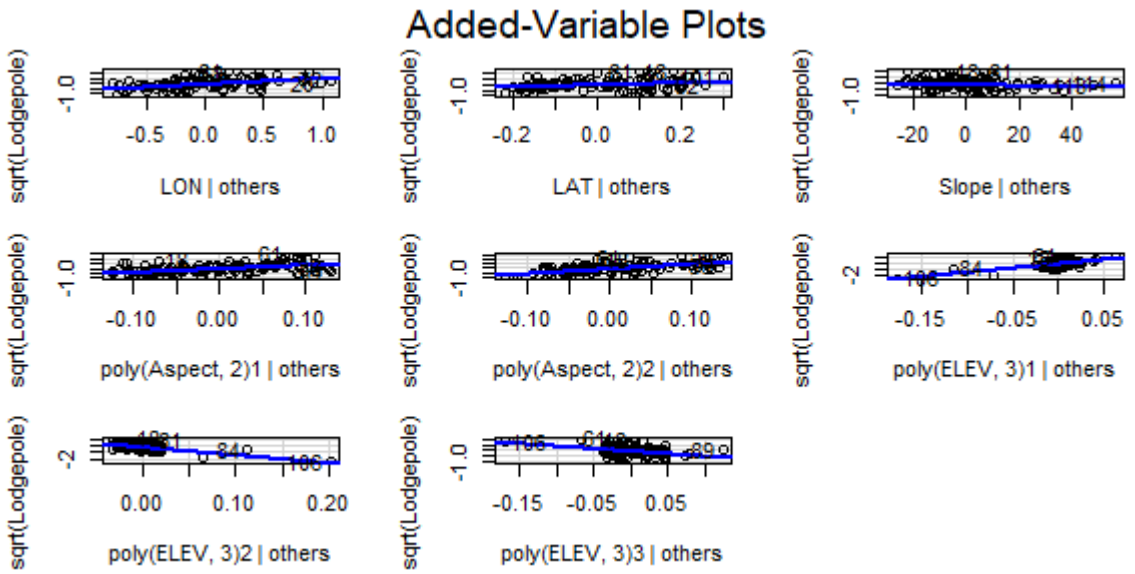

Fig. 8. Added variable plots of all covariates included in the model. Each plot looks linear, therefore we can conclude that linearity is met.

To test normality for this model, a KS test was performed to test the standardized and de-correlated residuals against a standard normal distribution. The p-value from that test was 0.1352, which is greater than 0.05, indicating that we can assume normality. The distribution of the standardized residuals can be seen in the histogram in Figure 9. This histogram seems to be normally distributed, thus this assumption is met.

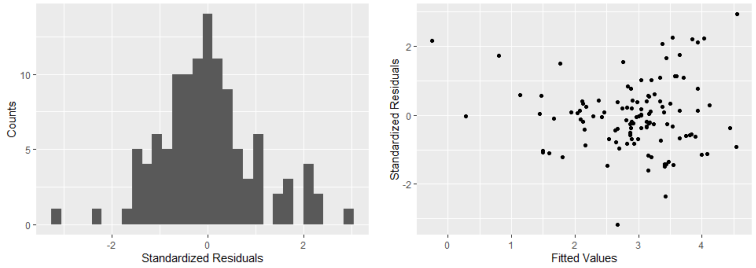

Fig. 9. Histogram of standardized and de-correlated residuals and scatter plot of fitted values compared to standardized and de-correlated residuals. These residuals appear to be approximately normal, and the variance of these residuals seems to be fairly constant.

The validity of the equal variance assumption can be seen in the scatter plot in Figure 9 which compares fitted values to standardized and de-correlated residuals. This graph seems to have points scattered evenly across the space with no obvious pattern, thus I will assume that equal variance is met.

The model also depends on the assumption that the lodgepole basal area for one tree is independent from the other trees. Though this assumption is not true, I accounted for this by using a spherical correlation structure in the model. Figure 10 shows the correlation after incorporating this correlation structure into the model. The red dots show the correlation after fitting the described model. Although, it appears that there may still be some correlation unaccounted for, I feel comfortable proceeding with this analysis because the red points level off on the right side of the graph, but the black dots increase, indicating the correlation has been accounted for.

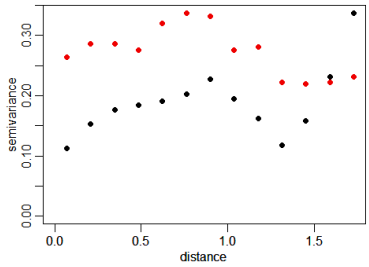

Fig. 10. Variogram after adding spherical correlation structure to the model is shown in red with the original correlation shown in black. It appears that there may be some correlation still not accounted for, but the red dots flatten out on the right side of the graph while the black dots continue to increase. Thus, including the spherical correlation structure does account for the correlation in the data.

4 Model Fit and Predictive Ability

To determine how well the covariates included in the model explain square root lodgepole basal area, I calculated a psuedo-R^2 value, which is 0.798. This value is difficult to interpret, but because the value is somewhat close to 1, I believe that this model does an good job of explaining lodgepole basal area. Because this value isn't closer to 1, perhaps additional covariates could be added to the model which would provide more understanding to what affects the lodgepole basal area.

To assess the predictive ability of the model, a cross validation study was conducted to calculate RPMSE (root predictive mean squared error), bias, average prediction interval width, and average prediction interval coverage with 1000 repetitions. The RPMSE from the cross validation study was 3.41, meaning that the predicted lodgepole basal area for a tree was off from the true value by 3.41 on average. This is very low when compared to the standard deviation of the original logepole basal area, which is 6.24. The average bias was 0.23, which which is essentially 0, so there does not appear to be any concerning bias in the predictions. The average coverage was 0.91, meaning that the 95% prediction intervals contained the true value 91% of the time. This is pretty good, but I would prefer it to be closer to 95%. The average interval width was 9.86, which is slightly larger than one standard deviation of the data, so this shows that our model predicts fairly well.

Based on these diagnostics, the predictive ability of this model appears to be fairly strong. Even though the prediction intervals are a little wide compared to the standard deviation of the response variable, the range of lodgepole basal area is about 34.61, so the intervals still provide some information.

5 Results

After fitting this model to the data, I obtained coefficients and 95% confidence intervals as shown in Table 2. It is important to remember that these coefficients explain the affect of the covariates on the square-root of lodgepole basal area. However, after testing a model which didn't use a square root transformation on lodgepole basal area, I saw the same signs on each coefficient as I see in this chosen model. Thus, the direction (positive or negative) of the effect of each covariate on lodgepole basal area is the same direction as the effect of each covariate on square root lodgepole basal area.

Coefficient	95% LB	Estimate	95% UB
Intercept	-116.85	-8.84	99.17
Longitude	-0.48	0.2	0.88
Latitude	-0.87	0.84	2.55
Slope*	-0.011	-0.007	-0.003
Aspect*	1.9	2.56	3.21
Aspect^2*	2.48	3.24	4.0
Elevation*	2.42	3.33	4.25
Elevation^2*	-7.32	-6.58	-5.84
Elevation^3*	-2.34	-1.65	-0.96

Table 2. Coefficients and 95% confidence interval for each coefficient in the model. It is important to note that the (*) indicates which covariates were found to be significant in the model due to a significant p-value (< 0.05) and each confidence interval did not contain 0.

From these results, we can see that slope, aspect, and elevation all had a significant effect on square root lodgepole basal area measurements. We can see that slope has a negative influence on square root lodgepole basal area. Additionally, it appears that elevation and aspect have a positive influence on square root lodgepole basal area. Each of these coefficients can be interpreted in a similar way, for example, we can interpret the coefficient for slope as follows: As slope increases by 1 unit, the square root lodgepole basal area will decrease by between 0.011 and 0.003 , on average.

Based on these results, I would conclude that trees with high aspect, high elevation and low slope, tend to have larger basal areas. This makes sense because I imagine that an area with a steep slope may not have as good of foundation for tree roots to thrive. Additionally, it could be that there are fewer pine beetles in areas of high elevation, leading to more tree growth. I would recommend that new trees be planted in areas with high elevation and high aspect and low slope.

With this model, I can now predict lodgepole basal areas for trees that the FIA were not able to travel to. These predictions are shown in Figure 11 below. It is important to note, that these predictions are on the original scale of the data. From this graph, it appears that there is a certain region in the forest that is best suited for lodgepole pine trees. This area is shown by the red and yellow dots. I am curious to look into this further. My guess is that this area is indicating a peak in the forest where the trees are higher and perhaps have a high aspect. Essentially, I feel satisfied that this model predicts fairly accurately for unknown measurements.

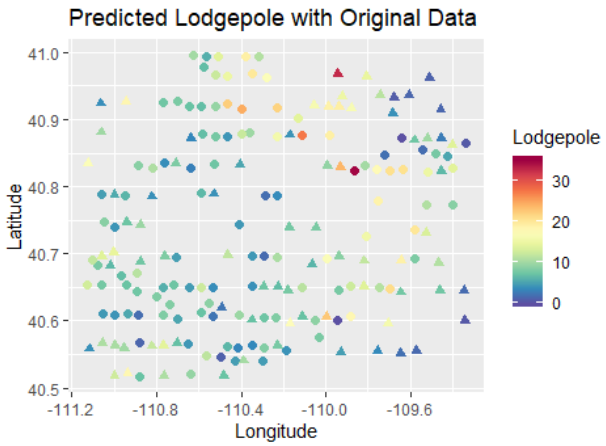

Fig. 11. Predictions for missing trees in the dataset. Predictions are shown with a triangle and original values are shown with a circle. There appears to be a certain area where lodgepole pines thrive the best, as shown by the yellow to red dots. However, for most of the region, lodgepole pine trees seem to have small basal areas.

6 Conclusion

From these results, I conclude that the best environment for lodgepole pine trees to grow is an area with low slope and high elevation and high aspect. Additionally, I can use this model to predict for trees that are not able to be measured. Although these predictions seem fairly accurate, the coverage is a little low and the prediction interval width calculated from my model seemed a little large. However, these results seem good enough and all the diagnostics looked good, so I feel satisfied using this model to predict for additional trees.

One shortcoming of this model is the transformations I performed on the data to justify linearity. Perhaps, there is a better transformation that would have had even better diagnostics and perhaps would have accounted for the non-linearity more. I'm especially interested in studying additional transformations on the aspect variable because of its unique circular nature. Perhaps, a different transformation would account for this better. Further, the variogram after fitting the model showed that there might still be some correlation that is unaccounted for in the model since the red dots were not very close to 0. Further testing could be done to try to account for more correlation in the data.

I believe one way to improve this model would be to add additional covariates that could explain more about what affects lodgepole basal area. These covariates could include information about the type of soil or amount of precipitation that area receives, although this information may be difficult to collect.