MIDTERM: LODGEPOLE PINE BASAL AREA

Jeremy Meyer

1 Introduction

Due to warmer conditions in recent years, pine beetles have been negatively impacting Lodgepole pines in Unita National Forest (located in Utah, USA). In the past, frigid winter temperatures have killed off enough beetles to keep the environment in an equilibrium. Due to the warmer winters recently, pine beetles have been having a heyday of killing pine trees and hindering their growth. However, not all trees in the Unita have been damaged severely. The FIA (Forest Industry Analysis) visited the forest and collected data on several Lodgepole trees scattered around the area. We wish to study the effects certain environments have on pine tree growth. This could help foresters better understand the decline in Lodgepole growth and help with sustainable forest management. The goals of this analysis are to address (1) what environments are conducive to Lodgepole pine tree growth and (2) what the tree growth is like in areas the FIA was not able to visit.

1.1 The data

The FIA sampled several small plots in the forest for the data. In order to measure Lodgepole pine health/growth, they added the total basal areas of each tree in the plot. This was called the cumulative Lodgepole basal area (ft^2/acre), which we will refer to as Lodgepole area. They also measured the latitude, longitude, elevation (in ft), average slope of the plot, and slope aspect (direction from due north slope is facing). Degrees were used to measure latitude, longitude, the slope (0=flat, 90=vertical) and aspect (0=slope facing North, 90=West, 180=South and 270=East). We will explore how each of these environmental factors affects pine tree growth. The data can be seen in Figure 1 and Figure 2.

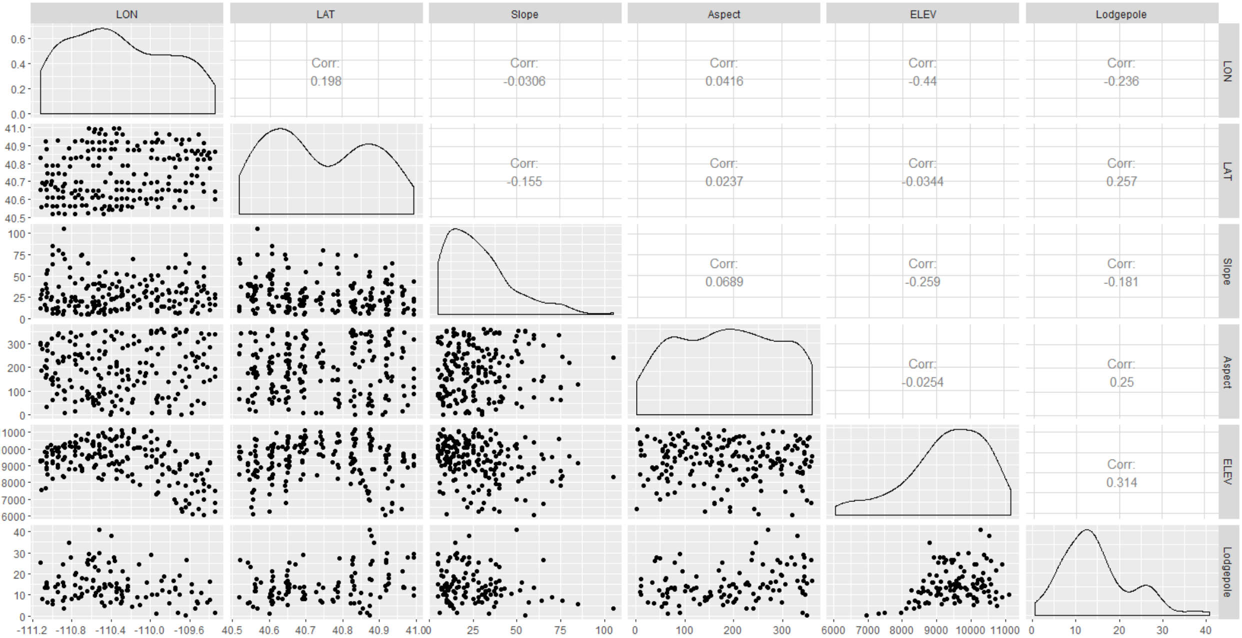

Figure 1: Lodgepole Area (Lodgepole), the environmental variables, and their correlations.

The dataset also contains 78 locations where the FIA was missing measurements of Lodgepole area. However, since Lodgepole area is the variable of interest, imputing these values gives no new information about the relationship

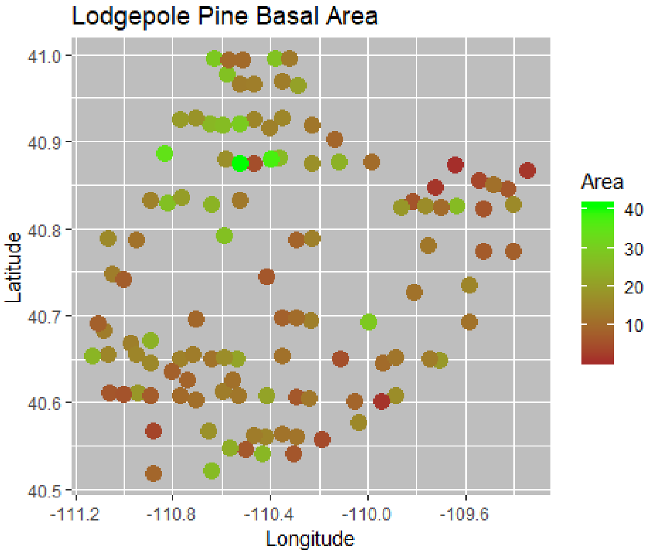

Figure 2: Lodgepole Basal area measurements by location. Note that observations closer in space tend to be more similar

between environmental variables and Lodgepole area. Instead, we will use the missing locations after building a model. We will then predict the Lodgepole area for the missing measurements and thus answer the second research question. As long as the reason the reason the FIA did not record basal area in the missing locations is not somehow related to Lodgepole area, this will not add bias to our results.

If we are to do some type of regression on the data, we must take care of the aspect variable, which is a circular quantitative variable. Since 1 and 359 degrees both correspond to similar values (close to due north), we will have to do some transformation to allow aspect to be included meaningfully in the model. The response variable (Lodgepole area) is strictly non-negative. Since some of the basal areas are very close to zero (values were as low as 0.47), we want to protect against predicting nonsensical negative values. Thus, we will take the log transform of Lodgepole area before our analysis. Lastly, from Figure 2, it can be seen that observations closer to each other tend to be more similar (or correlated) in space. This dependence in space must be accounted for in the final model.

2 The Model

In order to take into account the spatial dependence mentioned earlier, consider a spatial statistics model that follows an exponential spatial correlation structure. That is, we will assume the data follows a Gaussian process:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1(\mathbf{s}_1) \\ Y_2(\mathbf{s}_2) \\ \cdot \\ \cdot \\ \cdot \\ Y_n(\mathbf{s}_n) \end{bmatrix} \sim N\Big(\boldsymbol{X\beta}, \sigma^2((1-\omega)\boldsymbol{R} + \omega\boldsymbol{I})\Big) \tag{1}$$

Where $\mathbf{Y}$ is the vector of all observed, log-transformed basal area values, with each individual log Lodgepole area (Y_i) observed at spatial locations $\mathbf{s}_i$. We will assume that Lodgepole area follows a multivariate normal distribution with a mean of $\boldsymbol{X\beta}$, where $\mathbf{X}$ is a design matrix with a column of 1s for the intercept and numeric columns for the explanatory variables (details on this in section 3.1). $\boldsymbol{\beta}$ is a vector of coefficients that represent the "effects" of each of the explanatory variables included in $\mathbf{X}$. The term ω signifies a nugget effect, which allows some variation to exist at points in identical points in space. This is realistic because we would expect several measurements taken in identical places to not all be exactly the same.

This model assumes some dependence between the basal areas across space. Using an exponential spatial correlation structure, the correlation matrix $\boldsymbol{R}$ at row i and column j is defined as:

$$\mathbf{R}_{i,j} = exp\left(\frac{-\|\boldsymbol{s_i} - \boldsymbol{s}_j\|}{\phi}\right) \tag{2}$$

Where ϕ represents the range parameter, or a measure of how far the correlation across space lasts. The double bars indicate euclidean distance between the spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$. For this data, we will use latitude and longitude as spatial covariates. Then, to create a covariance matrix, we will multiply the correlation structure by σ^2, an estimate of the overall variance.

The exponential structure was used because it creates greater correlation between measurements that are close in space and gradually fades as measurements get further away. It also does not require the Lodgepole area measurements to be equidistant due to the continuous nature of the exponential function. This will allow us to capture the spatial dependence in the data and make informed predictions on areas the FIA was not able to go. Since we have also included the explanatory variables in $\boldsymbol{X\beta}$, we will also be able to answer questions about what kinds of environments are conducive to Lodgepole pine tree growth from the model by inference.

3 Model Justification

3.1 Variable Justification

Now we will consider which explanatory variables to include and how to structure our design matrix. Latitude and longitude are included as part of the correlation structure and slope and elevation are numeric variables that can simply be thrown into the design matrix. As mentioned earlier, aspect is a cyclical numeric variable and needs to be transformed. Consider breaking up the directional angle into its sine and cosine components on the unit circle and adding it to the X matrix. This will allow us to make inferences on slopes that face North-South (sine components) and East-West (cosine components). If these coefficients are negative, that means that slopes that face the West and South directions have a positive effect on Lodgepole area. It is worth noting that these values will be constrained such that $\sin^2(\text{aspect}) + \cos^2(\text{aspect}) = 1$, so we will lose some interpretability in the β coefficients. The aspect was also converted to radians before plugging it into the trigonometric functions. Recall that we log transformed the response variable, Lodgepole area, to enforce positive values and improve model fit. Thus our final model is:

$$\begin{aligned} log(\mathbf{Area}) \sim & \beta_0 + \beta_1(\sin(\frac{\pi}{180}\mathbf{Aspect})) + \beta_2(\cos(\frac{\pi}{180}\mathbf{Aspect})) + \beta_3(\mathbf{Elev}) + \beta_4(\mathbf{Slope}) + \boldsymbol{\epsilon}, \\ & \boldsymbol{\epsilon} \sim N(0, \sigma^2((1-\omega)\boldsymbol{R} + \omega\boldsymbol{I})) \end{aligned} \tag{3}$$

Or equivalently,

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2((1-\omega)\boldsymbol{R} + \omega\boldsymbol{I})) \tag{4}$$

Where the bolded variables indicate the corresponding data values for that variable arranged in a column, and ϵ indicates the error with the model. For future reference, $\sin(\frac{\pi}{180}\mathbf{Aspect})$ will be referred to Aspect.NS and $\cos(\frac{\pi}{180}\mathbf{Aspect})$ will be referred to as Aspect.EW.

3.2 Model Assumptions

Since we are assuming a linear relationship between log-Lodgepole area and the explanatory variables in Equation 3, we must check for linearity. Added-variable plots are shown in Figure 3 from the model in Equation (3). For the most part, the plots look fine, although are a couple outliers (observations 108 and 44) that may be worth looking into. There may be something unique about those locations.

Next we will check for residual independence, however, since the model has a non-diagonal covariance matrix, the residuals will certainly not be independent. Instead, we will check to see if the model accounts for all dependence after decorrelating the data. If we pre-multiply Equation 4 by the inverse of the lower cholesky decomposition of $(1-\omega)\boldsymbol{R} + \omega\boldsymbol{I}$, which we will refer to as $\boldsymbol{L}^{-1}$, we obtain the following equation.

$$L^{-1}\mathbf{Y} \sim N(L^{-1}\mathbf{X}\beta, \sigma^2\mathbf{I}) \tag{5}$$

Since the covariance is now a diagonal matrix, the residuals should be independent. After taking out the correlation structure, there shouldn't be any dependence between the residuals left. A map of the residuals across space is shown in Figure 4. We are looking for no patterns in the residuals. Since the positive and negative residuals look relatively mixed throughout the map, we can assume the model accounts for the correlation in the data.

Next, we will check normality of the decorrelated residuals. According to the histogram and QQ plot in Figure 5, it appears that because of a couple outliers, the left tail seems heavy. This won't necessarily cause too many problems in the analysis, but it might be worth looking into why one observation had a standardized residual of -3.95.

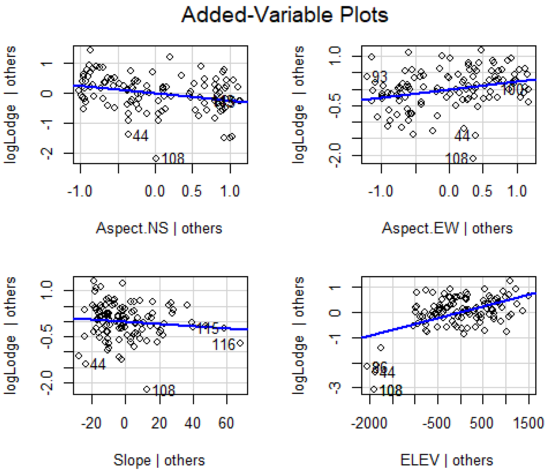

Figure 3: Added variable plots for the covariates in the spatial model.

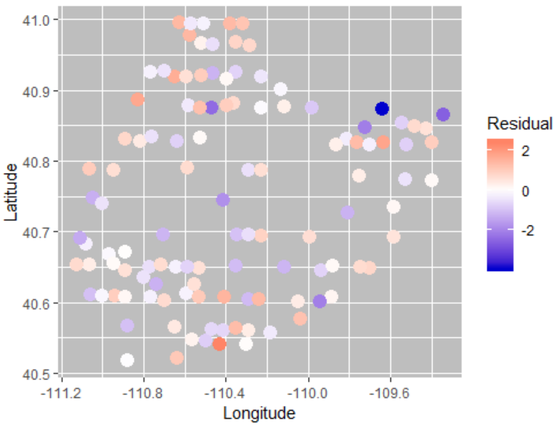

Figure 4: Residual map. We are looking for an absence of clusters between high and low residuals.

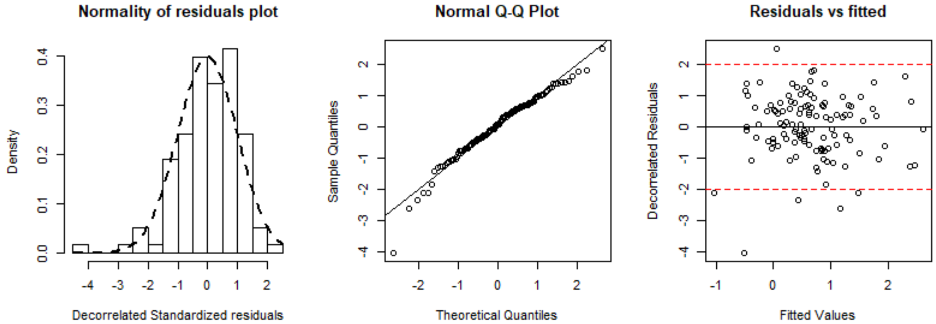

Figure 5: Histogram of residuals (left) with Normal QQ plot (middle). The left tail is a little heavy, so there are potential outliers. For checking equal variance (right), Decorrelated residuals plotted against fitted values for $\boldsymbol{L}^{-1}Y$

Finally, we need to check that the decorrelated residuals have equal variance. This can be seen in the right plot of Figure 5. Note that most of the residuals are contained within 2 standard deviations, with the exception of the outliers mentioned earlier.

3.3 Model Performance (Fit and Prediction)

Next we will evaluate model fit. After decorrelating the data, the R^2 value was 0.7277, which means that after accounting for correlation, the model accounts for 72.77% of the variation around its mean. To evaluate model prediction, we performed leave one out cross validation on the complete lodgepole data. For each iteration, we computed 3 metrics: prediction interval width, coverage, and root mean squared error.

Since the model predicts on the log scale, the predictions were first transformed back on their original scale before these metrics were calculated. This way the metrics are meaningful when compared to the original data. The average prediction interval width was about 35.3 (2.13 on log scale) with a coverage of 94.8%. These intervals are rather wide and cover about 87% (48% on log scale) of the total range of the data. More spatial data may be needed to help narrow the range of the intervals. The root predicted MSE on the original scale was about 7.144 (.553 on the log scale), which means the standard deviation of the model's prediction error is about 7.144. If we had used a Gaussian spatial correlation structure instead of an exponential, we would've gotten a slightly higher rpMSE of 7.248.

4 Results

We will address (1) what environments are conducive to Lodgepole pine tree growth and (2) what the tree growth like in areas the FIA was not able to visit.

To answer (1), we will examine the β coefficients from the model located in table 1. The β coefficients are interpreted as an expected increase for every 1 unit increase in log(Area) holding all other variables constant. However since the Aspects are constrained, we can only interpret the signs since we cannot change one without holding the other constant. Based on the table, we would conclude that slopes facing South and East have a positive effect on Lodgepole area since 0 is not contained in the confidence interval. Elevation seems to have a positive effect on Lodgepole area. Since the response is on the log scale, for every foot increase in elevation, we would expect on average exp(.0005)=1.004 multiplier effect or 0.4% increase in Lodgepole area holding all other variables constant. For a thousand foot increase in elevation, the multiplier effect is exp(1000*.005) = 1.59. A similar interpretation holds for slope, but we would not conclude that steepness of slope significantly affects Lodgepole area since 0 is contained in the 95% confidence interval.

The range parameter (ϕ) is a measure of how long the correlation lasts as the distance increases. THe effective range, or distance required to generate a correlation of 0.05 is a distance of -log(.05)*.2213 = 0.662 degrees. However, this is difficult to interpret or convert to miles since the distance between longitude degrees changes depending on the latitude. The nugget effect value of 0.6077 indicates some variation of measurements at identical points in space and σ^2 is a measurement of the error the model cannot explain.

Table 1: Model Coefficients and Confidence Intervals

	Int (β_0)	Asp.NS (β_1)	Asp.EW (β_2)	Elev (β_3)	Slope (β_4)	ϕ	ω	σ^2
2.5%	-3.3798	-0.4056	0.1379	0.0003	-0.0097	0.0339	0.2665	0.2189
Estimate	-1.8799	-0.2698	0.2891	0.0005	-0.0036	0.2213	0.6077	0.3382
97.5%	-0.3800	-0.1341	0.4402	0.0006	0.0025	1.4455	0.8685	0.5226

For the second research question, we will use the model to predict at the locations with missing Basal Area values. The results are shown visually in Figure 6 and a table of predictions with uncertainties are shown in Table 2.

Table 2: Sample Predictions for 4 Points in Top Left of Figure 6

LON	LAT	Slope	Aspect	Elev	Prediction	2.5%	97.5%
-111.123	40.834	12	303	9309	19.282	6.227	59.712
-110.944	40.927	20	335	8909	19.272	6.230	59.619
-111.062	40.881	35	230	8769	10.340	3.322	32.189
-111.064	40.924	40	158	7673	4.118	1.260	13.463

In summary, elevation and slope aspects facing south and east had a significant positive impact on lodgepole area. For every 1000 foot increase in elevation, we expected on average a 59% increase in Lodgepole area. Slope was not found to be significant. We were also able to predict what Lodgepole areas were like in areas that had missing measurements in Figure 6.

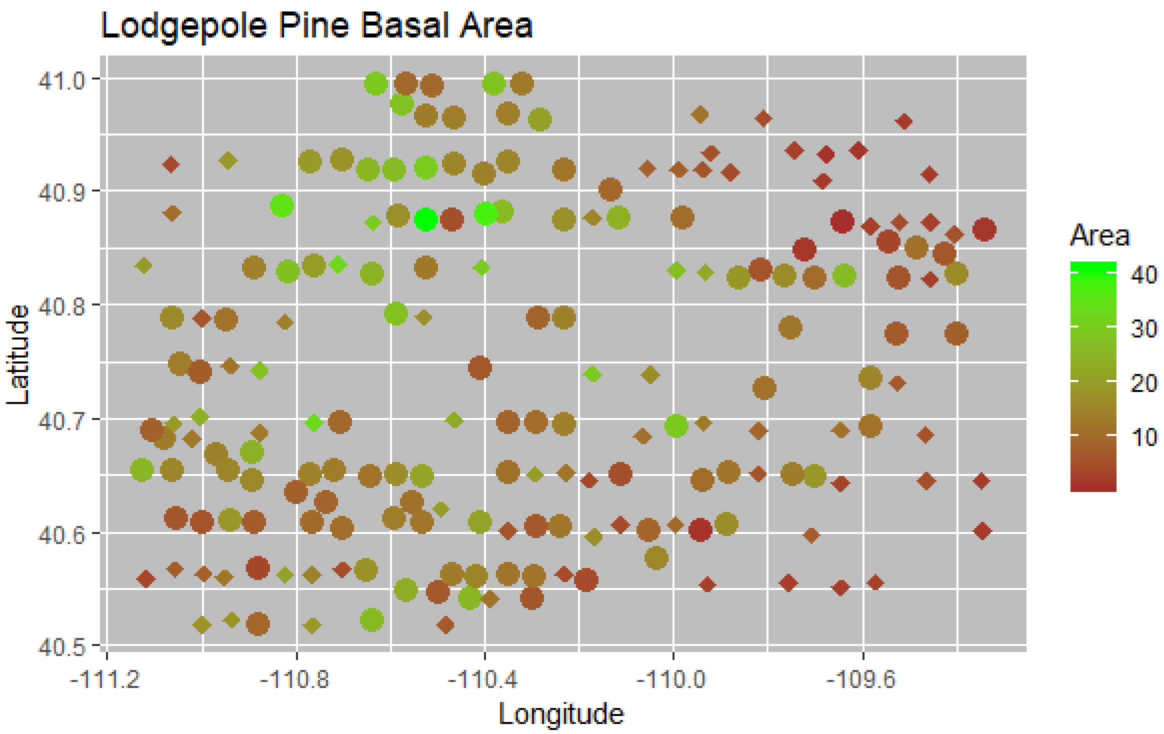

Figure 6: Map of observed Lodgepole areas (circles) and predicted (diamonds) areas. A table of predictions for the 4 predicted points in the top left of the map are shown in Table 2

5 Conclusion

We answered the goals of the study by using an exponential spatial correlation structure with a Gaussian process. This allowed us to not only figure out how Lodgepole area was affected by its environment, but it allowed us to capture the spatial dependence in the data. We found that elevation had a positive effect on pine tree growth, which is likely due to the cooler temperatures that kill off beetle populations. We also found the southern and eastern slope aspects had a positive effect on growth and made a map of Lodgepole area predictions in areas the FIA was not able to visit.

One major shortcoming of the model is that after transforming area to its original scale, the prediction intervals are very wide. They tend to get very wide around the upper bounds (see Table 2), especially for higher predictions. This may not be particularly useful if they cover the majority of the range of the data. There were also a couple outliers that the model can not explain very well. There may be some other variables at play that are driving those few outliers. Another limitation of this study is that it does not take into account how the growth changes over time. Making predictions with this model will likely only work for a few years if Lodepole pine growth keeps declining.

One next step we could take is to examine the outliers and try and examine why those particular locations did not behave like most other places. It may also be worth exploring new locations to narrow our prediction intervals and impute the missing Lodgepole area values with true values. Although this analysis is helpful in determining how certain locational factors affect growth, connecting these factors to pine beetles may be really helpful for foresters. For example, since the slopes facing north tended to have a negative effect on Lodgepole area, studies could be done to see if those conditions facilitate higher pine beetle populations.