Analysis of Lodgepole Pine Growth

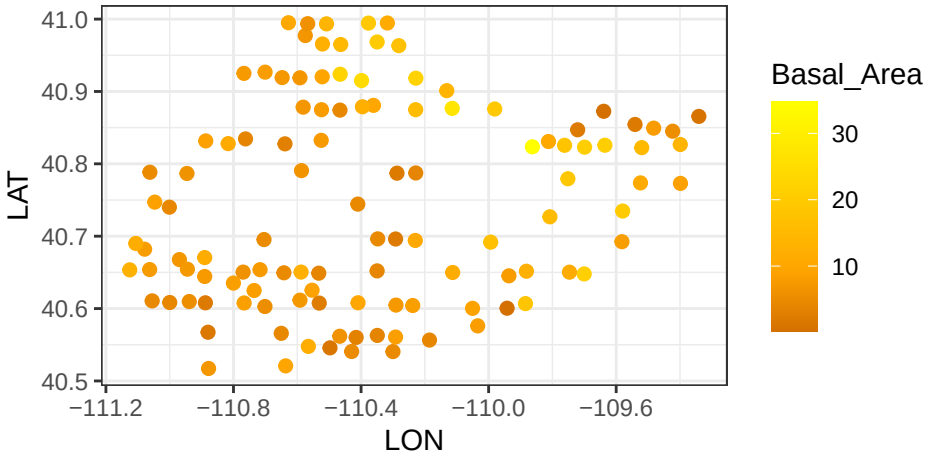
Carly Lundgreen

March 16, 2021

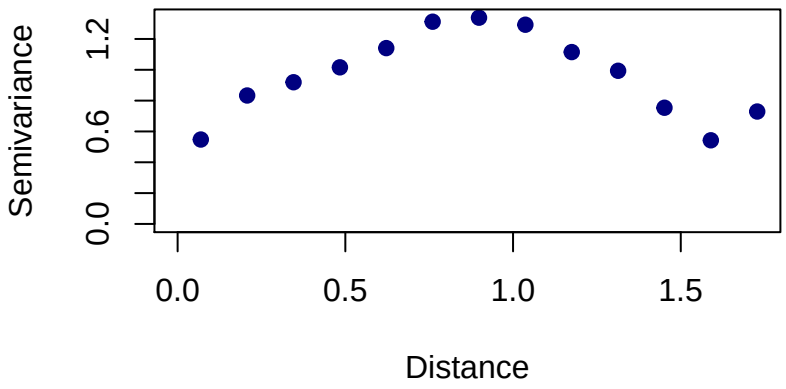
Introduction

The Uinta Mountain Range and National Forest in Utah is home to a wide array of tree species. In recent years, wildfires and the Mountain Pine Beetle have caused significant damage to the forest. As part of the Forest Inventory Analysis (FIA), researchers are interested in understanding the health of the Uinta National Forest by evaluating the growth of trees in the area. The Lodgepole Pine is a particular tree species that, in Utah, is only found in the Uintas, and the health/growth of lodgepole lines can be used as an indicator of overall forest health. As a measure of tree growth, researchers measured the basal area of lodgepole pines at various locations throughout the Uintas. Additional environmental information–such as coordinate location, steepness of the mountain slope, aspect (compass direction of the slope), and elevation– was recorded at the time of basal area measurement. The aims of this analysis are to understand how basal area is affected by the environment (slope, aspect, and elevation), explain the types of environments that are conducive to lodgepole pine growth, and predict basal area (a measure of tree size) for trees at locations where the FIA was unable to go.

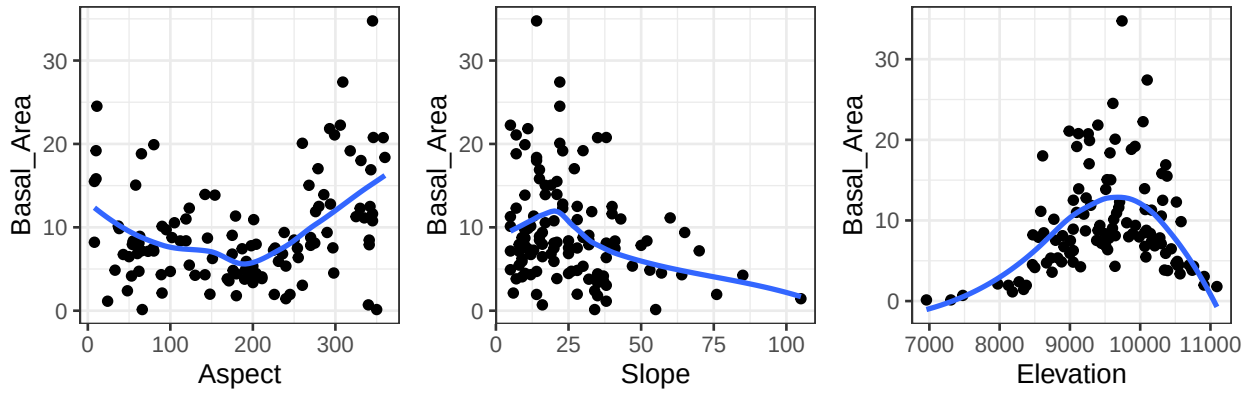
This dataset had 192 rows with 78 missing basal area measurements. The operating dataset was created by including only the non-missing obserations and the statistical model that was fit to this operating dataset allowed us to impute the 78 missing basal area measurements. Latitude, Longitude, Slope, Aspect, and Elevation were all recorded by researchers as covariates in this dataset. The plot below illustrates the locations of each of the 114 basal area measurements. It appears that large basal area measurements tend to be near other large basal area measurements. This is most obvious near the (-110,40.9) coordinate location. These similarities in basal areas denote the likely presence of spatial correlation between these measures of tree size. This will be discussed in more detail in further sections.



The (semi)variogram below shows that the variance of observations tends to increase as distance increases, which is indicative of spatial correlation present in these data. This will be accounted for in the statistical model that will be outlined in following sections. If spatial correlation is not accounted for, the standard errors associated with inference procedures will be inaccurate, meaning inferential conclusions will be invalidated. Prediction accuracy will also be decreased because we are ignoring an important component the distribution of the response variable.



The plots below illustrate the relationship between the response, basal area, and the covariates that are of interest to the researchers. Nonlinear relationships are obviously present between these variables and the response. For example, as aspect increases, basal area appears to decrease first and then begin to increase, on average. We will account for this nonlinearity in our statistical model in order to avoid inaccuracies in predictions.



Model Specification

In order to unerstand the relationship between lodgepole pine growth and these environmental factors (as well as generate predictions of basal area), we will fit a generalized least squares model with an exponential correlation structure and variance nugget. Within the regression "portion" of the model, we have accounted for the nonlinearity among aspect, slope, and elevation through the use of smoothing splines. The model is outlined below:

$$y_i = \beta_0 + s(x_{Aspect}) + s(x_{Slope}) + s(x_{Elevation}) + \epsilon,$$
$$\epsilon \sim N(0, \sigma^2((1-w)R + wI))$$

Where y_i represents the i^{th} basal area measurement and β_0 indicates the expected average basal area for a tree with zero aspect, zero slope, and zero elevation. This does not have valuable interpretation in the

context of our analysis, so we will not report an estimate of β_0. Note that s in the above model indicates the application of smoothing splines to each of the covariates. The smoothing spline process aims to identify a smooth function $g(x_i)$, where x_i is either aspect, slope, or elevation, where the quantity $\sum_{i=1}^{n}(y_i - g(x_i))^2$ is small. Specifically, the smoothing spline identifies the function that minimizes the penalized residual sum of squares:

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 - \lambda \int g''(x)^2 dx$$

Where λ is a "smoothness parameter" that penalizes the second derivative of g so that the function is not overly erratic to avoid overfitting to the data. The ideal smoothness parameter is selected through cross validation techniques. The solution to the above expression is that $g(x_i)$ is a natural cubic spline with knots at each x_i with the coefficients then shrunk back towards zero. The main advantage to smoothing splines is that it is not necessary to select knot points because each x_i is used as a knot. However, smoothing splines do not provide easily interpretable coefficients. Therefore, in order to understand the effects of aspect, slope, and elevation on basal area, we will evaluate whether the smoothing spline is significant (which would tell us how the covariates are behaving in relation to basal area). We will then evaluate univariate plots to determine these relationships.

In the model, $\mathbf{R}$ is an nxn matrix where the ij^{th} elements are based on $\rho(s_i, s_j)$, or the correlations between basal area measurements at differing locations. The elements of $\mathbf{R}$ are as follows:

$$R_{ij} = exp\frac{-||si - sj||}{\phi}$$

Where ϕ is a range parameter such that when ϕ increases, the spatial correlation between observations at locations i, j also increases. Note also that $||s_i - s_j||$ indicates the Euclidean distance between two basal area observations. The 'nugget effect' is indicated by w and allows for variability when $||s_i - s_j|| = 0$. If more than one observation is taken at a single location, it may not have perfect correlation with itself, especially if observations are taken years apart when the tree is likely to have grown between the two measurements. This nugget effect allows for sampling variability in these "same-distance" scenarios. The matrix $\mathbf{I}$ is an nxn identity matrix. Finally, the parameter σ^2 indicates the residual variance, or the variance between observations. Essentially, σ is the amount that we expect the model to differ from the true basal area measurements on average.

The assumptions of this model include equal variance of the residuals (denoted by the constant σ^2 parameter described above) as well as normality of the residuals. The validity of these assumptions will be explored in the following section.

Model Justification

Our model excluded the use of Latitude and Longitude as fixed effects due to the researchers' interests in understanding, specifically, the effect of the environment on basal area. The latitude and longitude effect is accounted for in the modeling of the spatial correlation.

When deciding which nonlinear function to apply to the aspect, slope, and elevation variables, a generalized additive model (GAM) approach was taken. A "full model" was fit by applying a smoothing spline to each of the three covariates and adding these functions together. Note that while deciding which variables to include and whether a smoothing spline was needed, we did not consider the spatial corrleation. This effect was added after variable/function selection took place. We then evaluated the significance of the variable and the significance of the smoothing spline through Likelihood Ratio tests. "Reduced models" that 1) excluded the variable in question entirely and 2) included the variable linearly rather than with a smoothing spline were also fit for each of the three covariates to then compare to the full model via a likelihood ratio test.

The p-values for each of these ANOVA tests was less than 0.001. This means that, in all cases, the "full model" was a better fit than the "reduced model". Therefore we kept all three covariates in the model with a smoothing spline.

After selecting the variables and applying smoothing splines, we used cross validation to determine the optimal smoothness parameter, λ, for each of the univariate fits between the covariates and basal area. These optimal values are included in the table below.

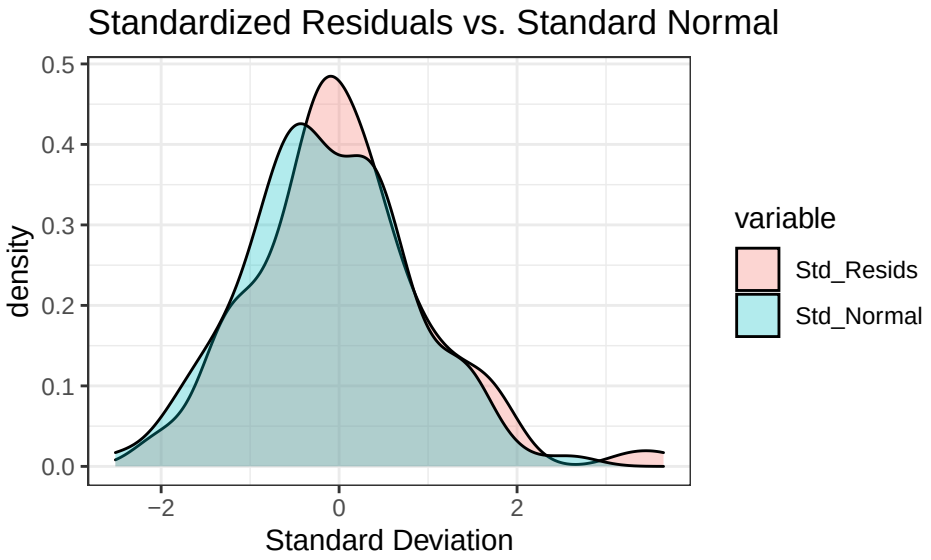
	Smoothing Parameter
Aspect	0.775
Slope	1.500
Elevation	0.993

The following table compares an ordinary least squares (OLS) fit that did not account for any spatial correlation to our selected model. AIC and BIC information criterions were reported due to the inference and prediction-focused goals of this analysis. The AIC and BIC values were both smaller for the GLS fit in comparison to OLS, which means that a model accounting for spatial correlation fits these data better.

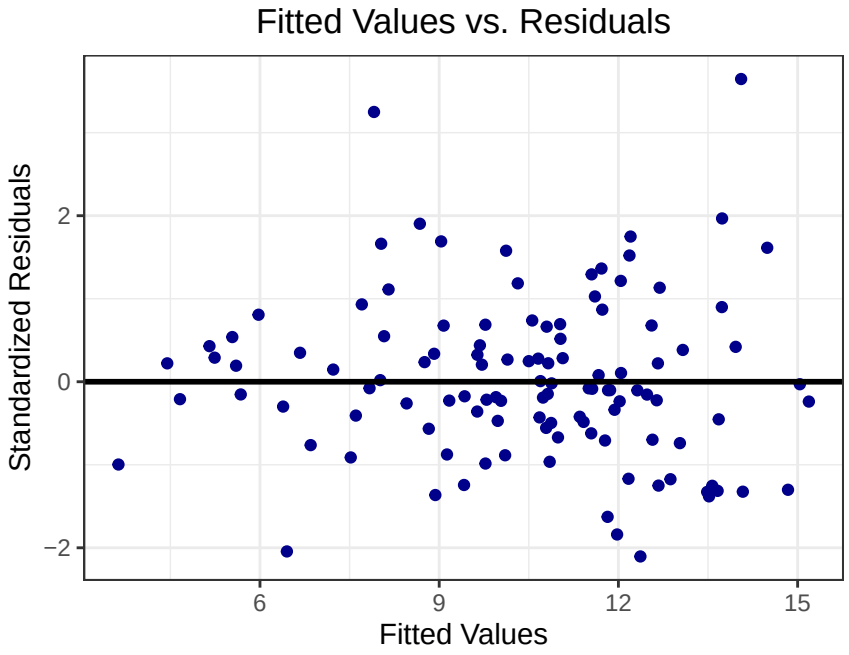
	AIC	BIC
OLS	756.0549	769.5573
GLS	715.6342	734.5376

Another metric that is indicative of model fit is a "Pseudo R^2" value, calculated as $R^2_{pseudo} = Cor(y, fitted(mod))^2$. This value can begin to give us an idea of the proportion of variation in basal area measurements that is explained by our model, according to the standard R^2 interpretation. Our calculated R^2_{pseudo} was `r2_pseudo`, which does seem low. However, we only were provided three variables to explain variation in the basal area of lodgepole pines. In reality, there are many other variables that could explain the basal area of these trees. These variables may include time since last wildfire in the area, presence of mountain pine beetle damage, or soil quality. Because we only had three of dozens (or more) variables that can potentially explain lodgepole pine basal area, this R^2_{pseudo} value is not overly concerning.

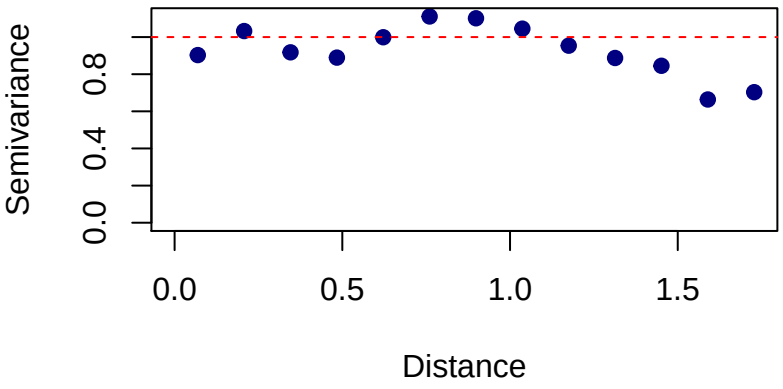
In order to evaluate the assumption of normality, we plotted a density curve of the standardized residuals from our model against the density curve of a standard normal random variable. It seems that there is a bit of a heavy right tail in comparison to the standard normal, a possible departure from normality, but the overal shape of the residuals is not a large cause for concern.



We also plotted the standardized residuals against the fitted values from our model to evaulate the assumption of equal variance. No overly concerning patterns in the residuals are seen in the plot below, although there does seem to be some residuals near fitted values of ten that are more closely grouped together than residuals from other locations. However, this grouping is not overly extreme, so we will say that the equal variance assumption is met.



We also want to confirm that the residuals from the model have been sufficiently "decorrelated." The semivariogram below shos that the variance of observations does not increase wildy with an increase in distance between observations. It varies in a consistent pattern around a straight line, with a slight decrease in very far distances. The residuals from our model appear to be sufficiently decorrelated.



Predictive Accuracy

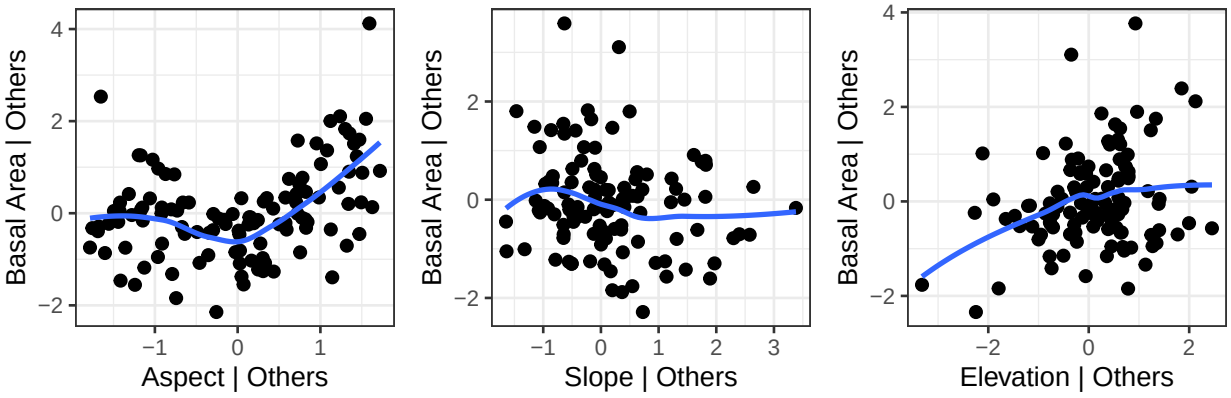
In order to assess predictive accuracy, we ran 50 cross validation studies on a training set of 20% of the data. We calculated the cross-validated root mean square error (RMSE), bias, average prediction interval coverage (i.e. what percent of prediction intervals actually contain the true value), and average prediction interval width for our model. These values are shown in the table below.

	RPMSE	Bias	Width	Coverage
	4.541	0.035	19.451	0.953

The RPMSE indicates we are off in our predictions of basal area, on average, by about 4.540944. This is about two thirds of the IQR of our response variable, which is 7.3425. This is perhaps a bit higher than we would prefer, but not overly concerning. The bias indicates that we are underpredicting, on average, by 0.0346975. This is very small relative to the range of the response variable. The coverage of 95% prediction intervals, is on average 0.9530435, which is as expected. The average prediction interval width, however, is quite wide in relation to the range of the response variable, which is 34.61. This would, ideally, be narrower, but our model still has decent predictive ability.

Results

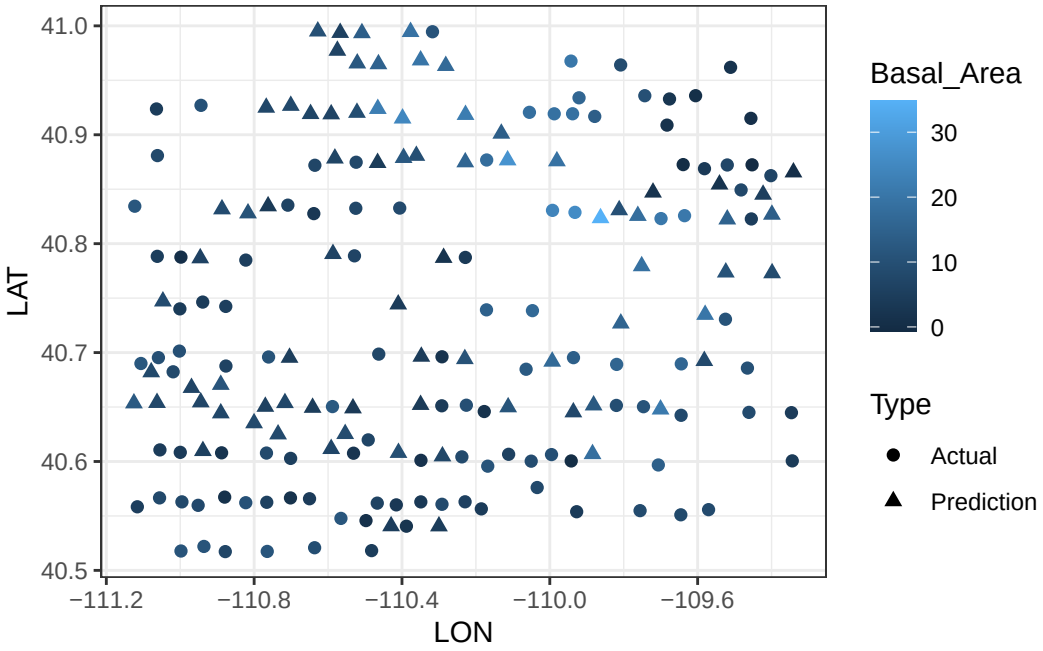
Because our model utilized smoothing splines, it lacks interpretable coefficients that can concisely explain the relationships between aspect, slope, elevation, and basal area. However, because we determined that smoothing splines were significantly necessary for each of the covariates in the model through likelihood ratio tests, we can evaluate the plots below to explain the relationship between our response and each of the covariates after adjusting for the other effects. Pseudo "added variable" plots were created, where each of the covariates and the response were adjusted for the other covariates and the residuals from these respective model fits plotted against each other.



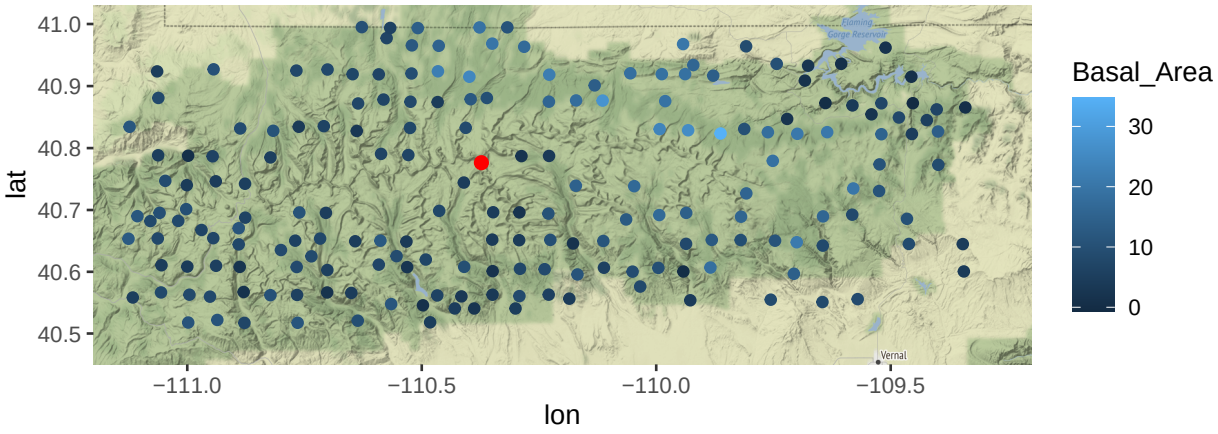
Based on the above plots, it appears that as aspect increases, basal area will, on average, decrease to a certain point and then begin to increase (on average). That shift appears to happen near the mean value of aspect, which was about 187. It doesn't appear that basal area, on average, changes very much as slope increases. If a relationship was to be gleaned, it seems that for small slopes, the average basal area is a bit higher than for very steep slopes. This is completely logical based on logic presented previously. It would not make sense for very large, and thus very heavy, trees to grow on extremely steep slopes. Finally, it seems that there is an average increase in basal area as elevation increases, although this increases seems to level off at very high elevations that are close to two standard deviations above average. This observation is

consistent with research by the Utah State University ecology department, which states that "the Lodgepole pine is adapted to high mountain slopes at elevations usually above 6,000 feet." This adaptation appears to be present in our dataset, based on these results. Based on our model, it seems that higher elevations, up to a point, and very large aspects are most conducive to lodgepole pine growth.

Using our model, we obtained predictions for areas that FIA was not able to measure lodgepole pine basal areas. These predictions are presented in the plot below along with the observed/actual basal area measurements. Based on this plot, it seems that we generally predicted large basal measurements near other large basal measurements, which is what we would expect due to the spatial correlation that we observed in these data.



We also plotted these predictions and observations on a geographic map that illustrates the terrain of the Uintas relative to the Basal Area values. It does appear that the larger basal areas (denoted by the light blue dots) are located along a mountain ridge, which is consistent with our conclusion that elevation is conducive to lodgepole pine growth. Also, the red dot below indicates the location of the highest point in Utah, King's Peak, and we do not see very large basal area messages near that peak. This supports our conclusion that the effect of elevation on basal area appears to level above a certain altitude.



Conclusion

By using smoothing splines to model the nonlinearity between aspect, slope, elevation, and basal area, as well as modeling the spatial correlation between tree basal area measurements using an exponential correlation function with variance nugget, we were able to better understand how the basal area of lodgepole pines is related to its environment via aspect, slope, and elevation. Similarly, we found that higher elevations appear to be conducive to lodgepole pine growth, until a certain point. Also, higher aspects are conducive to lodgepole pine growth. The relationship between slope and basal area was more difficult to ascertain.

The shortcomings of this model include the relatively few amount of covariates, which was discussed previously. Increasing the number of covariates used to explain basal area could provide a better model fit and potentially improve the "Pseudo R^2". Though our model had decent predictive ability, the prediction intervals were quite wide. Reduction of the standard error, potentially through more variables, would reduce the width of these intervals. Other functions could have been used to model the nonlinearity in these data, such as polynomial expansions, one-to-one transformations, or natural splines, and these functions may be better suited to these data. Further analysis of these data may include exploration of these alternative functions. Similarly, a Gaussian or Spherical correlation structure (as opposed to exponential) may have been worth exploring.

Overall, we have an idea of the conditions that are most conducive to lodgepole pine growth, and the predictions that we mapped above appear reasonable in relation to our inferential conclusions. A final note is that if researchers wish to understand overall forest health, perhaps other tree species are worth measuring apart from lodgepole pines. Currently, we can realistically only generalize these results to lodgepole pine growth, which may or may not be the best indicator of overall forest health in the Uintas.