Lodgepole Pine Basal Area Analysis

Josh Christensen

11/03/2021

Section 1: Introduction and Problem Background

The basal area of a forest is a good indicator of forest health. The U.S. Forest Service uses this metric to assess current status and overall trends within forests across the United States. The Forest Inventory and Analysis (FIA) program is charged with collecting this data, but it is impractical to try and gather data for all of the acreage that our forests cover. We would like to construct a model that can accurately predict basal area in locations where the FIA did not measure it, as well as provide insights into how environmental factors affect basal area. We will focus specifically on a dataset of lodgepole pine basal area from the Uinta National Forest in Utah.

Our data set includes the following variables:lodgepole pine basal area in ft.^{2}/acre, average slope of the plot in degrees, counterclockwise rotation from North in degrees, elevation of plot centroid in ft., and the geographic location of the plot in longitude and latitude.

We have displayed exploratory plots showing how each variable interacts with our response variable, lodgepole pine basal area.

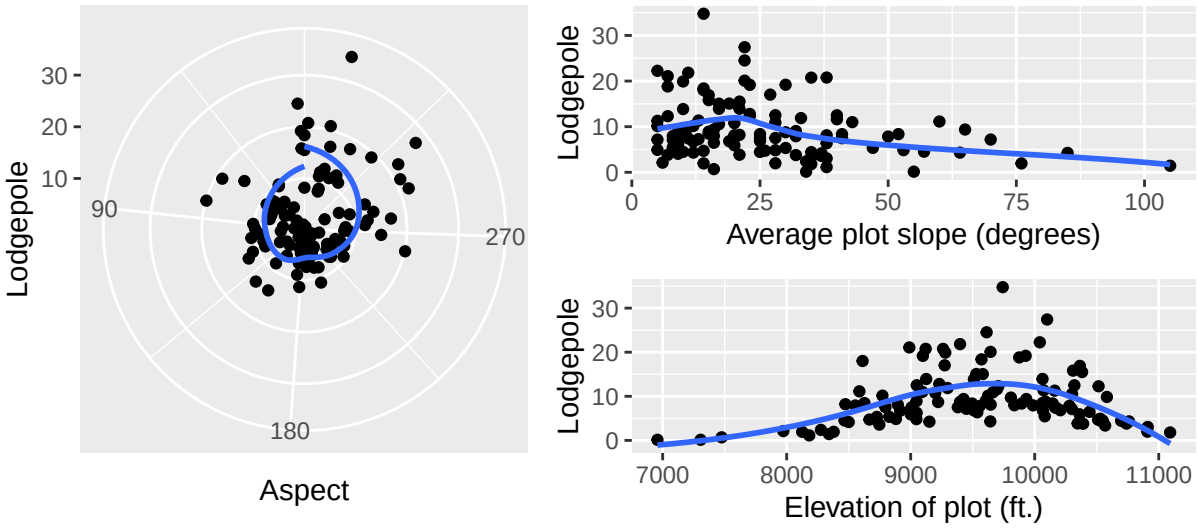

Figure 1: Relationships among variables

Our dataset has many potential issues. One is that there is spatial correlation between plot locations (i.e., plots that are closer together tend to be similar in basal area). This is shown in the variogram below. Another is the unequal variability in basal area as elevation changes. If we ignored these issues, our standard error calculations would be incorrect. This would in turn cause problems in our confidence intervals and prediction intervals for individual plots. We also note that we have a circular predictor in Aspect and an

apparent non-linear relationship between Lodegpole and elevation. If we ignored these data features our β estimates would not accurately represent the relationships between variables. By addressing these issues, we will improve our ability to make inference and to accurately portray the variation of the data, in addition to improving our predictive ability.

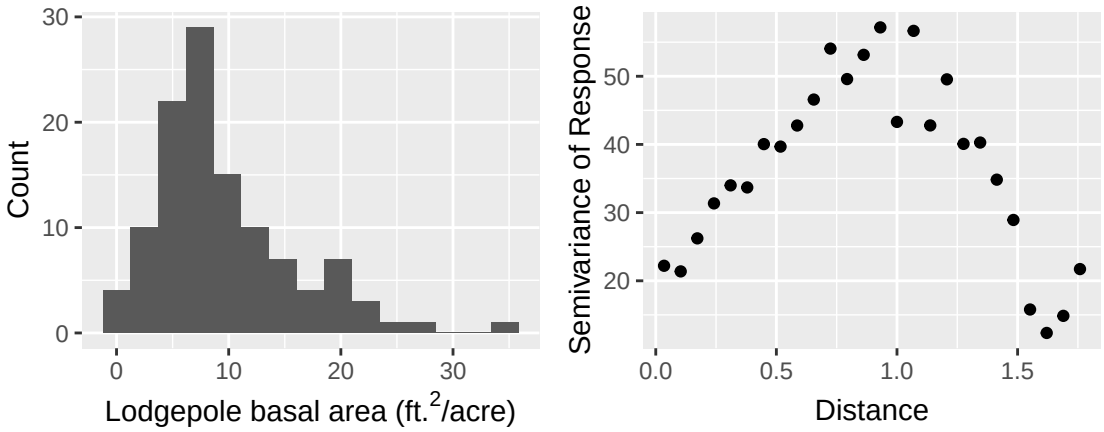

Figure 2: Response histogram and variogram

Section 2: Methods and Models

We will use a heteroskedastic spatial multiple linear regression model to analyze the lodgepole pine basal area data. We define the model for our data as $\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \sigma^2\Sigma)$ where $\boldsymbol{\Sigma} = \boldsymbol{DRD}$. In our model, $\boldsymbol{y}$ represents the lodgepole pine basal area of the plot. $\boldsymbol{X}$ represents the design matrix which includes our explanatory variables as well as a column of ones, representing the intercept. We included all three explanatory variables in our model, but with different features. We changed the Aspect variable into sin(Aspect) and cos(Aspect). This changes the circular relationship into two linear relationships. We also replaced elevation with a fourth degree orthogonal polynomial. We did this in attempt to capture the curvature evident in the relationship between elevation and the response. We chose a fourth degree polynomial because it outperformed other degrees in terms of AIC and BIC. Our parameters in this model are $\boldsymbol{\beta}$, σ^2, and the parameters of our covariance matrix: θ and ϕ. $\boldsymbol{\beta}$ is a vector of coefficients that quantifies the relationship between each respective explanatory variable and the response variable, in addition to setting the intercept. The effect of Aspect is split across the two $\beta_i's$ that correspond to the sine and cosine terms, while the effect of elevation is split across the four $\beta_i's$ that represent the polynomial. σ^2 represents the variance of the residuals from our model. The θ parameter quantifies the heteroskedasticity associated with elevation and defines, through the power variance function, the diagonal weights of the $\boldsymbol{D}$ matrix in our covariance matrix decomposition. ϕ defines the range of the spatial correlation. It is also used to calculate individual correlations using the gaussian correlation function within the $\boldsymbol{R}$ matrix in our covariance matrix decomposition.

Our model depends on four assumptions. The first is that all quantitative explanatory variables have a linear relationship with the response variable. The second is that the decorrelated residuals will no longer exhibit spatial correlation. While we do not assume spatial independence between plots, the validity of our model does depend on capturing the spatial correlation with our correlation function. The third assumption is that the residuals are normally distributed. The last assumption is that our $\boldsymbol{\Sigma}$ covariance matrix multiplied by σ^2 accurately represents the variability of the data. In other words, while our response is not assumed to have constant variance, we would like our covariance matrix to account for the changes in variance so that the underlying σ^2 is still constant. This is equivalent to constant variance of the residuals.

Section 3: Model Justification and Validation

We included all of the explanatory variables in our final model. All had significant effect on the response. We modified aspect into the sine and cosine of aspect so that our linearity assumption would be met. We modified elevation to a polynomial to account for curvature and made it orthogonal to keep our design matrix full rank.

We tested the linearity assumption using added-variable plots, which regress both the response and the explanatory variable against all other variables in the model and then display the resulting fitted values plotted against one another. This allows us to evaluate the relationship between the explanatory variable and the response, while accounting for the effects of all other variables in the model. The added-variable plots showed no clear deviations from a linear relationship.

We believe the assumption of independence except for spatial correlation to be reasonable. Intuitively, lodgepole pine basal area in one area should be independent of another area, provided the other area is across the forest from the first. Additionally, while we do not assume independence between plots that are close together, we do assume that the gaussian correlation structure captures the correlation. This is evident in the variogram of the decorrelated residuals shown below.

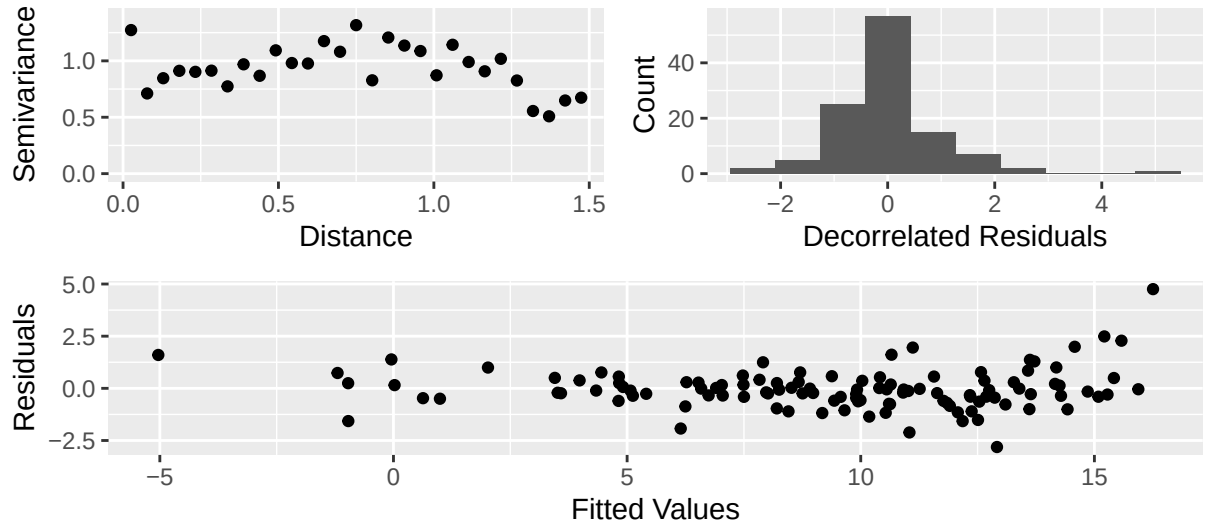

Figure 3: Assumption Plots

We checked the assumption of normally distributed residuals by creating a histogram of the standardized, decorrelated residuals. The histogram appears to show a normal shape, unimodal with most points within 3 standard deviations from 0. We do note that the outlier from EDA makes our plot less normal. We believe that our residuals are close enough to normal to satisfy this condition.

The final assumption of constant variance in our residuals was checked by plotting the standardized residuals against the fitted values from our model. The scatterplot of standardized residuals against the fitted values shows constant variance. Again we note that the outlier does not conform to this assumption as well as the rest of the data.

We calculated pseudo-R^2 as 0.59, indicating that roughly 59% of the variation in lodgepole pine basal area can be explained using the explanatory variables included in our model. The root mean square error(RMSE) is 4.002. This is moderately small relative to our data standard deviation of 6.24, which confirms that our model fits the data well.

We evaluate our predictions using root prediction mean square error (RPMSE), coverage, prediction

interval width and bias. We evaluate these metrics using Monte Carlo cross-validation. The average of each metric is reported in the table below.

Table 1: Prediction Diagnostics

	RPMSE	Bias	Coverage	Width
GLS	3.200	-0.003	0.965	16.447
Neural net	3.914	0.023	NA	NA
Bagging	3.928	0.280	NA	NA
Random Forest	3.913	0.253	NA	NA
Boosting	3.575	0.068	NA	NA

Given our negligible bias and relatively low RPMSE we find our predictions to be accurate. We also note that our gls model outperformed various machine learning algorithms in terms of prediction. Our coverage is also very close to the expected 0.95. Our prediction interval width indicates that we will generally construct intervals with a margin of error smaller than 8.22. This indicates that we can predict lodgepole pine basal area with confidence.

Section 4: Results

Since our model proved to be a good fit of the data, we were able to determine which environmental factors increase a plot's lodgepole basal area on average. Flat plots have more basal area on average than sloped plots. Basal area decreases by about 0.05 for every extra degree of slope (0.01,0.08).The effects of aspect and elevation are less obvious from their coefficients due to the modifications we have made to the variables prior to regression. For clarity on the true effects of elevation and aspect we have included plots indicating the partial effect of each. From these plots we see that plots facing north and northeast have more basal area on average (maximum at 331.2°). We also see that basal area is largest in a range of elevation between 9,500 ft. and 10,000 ft. (maximum around 9,645 ft.). θ represents how variance increases with elevation and ϕ represents the range of spatial correlation.

Table 2: Coefficient Table

	Estimate	Lower	Upper
Intercept	10.42	4.39	16.44
Slope	-0.05	-0.08	-0.01
sin(Aspect)	-1.70	-2.46	-0.95
cos(Aspect)	3.12	2.30	3.94
Elevation	16.41	7.12	25.70
Elevation^2	-30.85	-36.56	-25.14
Elevation^3	-12.50	-18.11	-6.90
Elevation^4	6.09	0.60	11.57
θ	0.86	-0.12	1.84
ϕ	0.59	0.32	1.11

We used our model to predict the lodgepole basal area of all of the plots in our dataset that did not have the basal area recorded. These predictions are displayed in the heat map below.

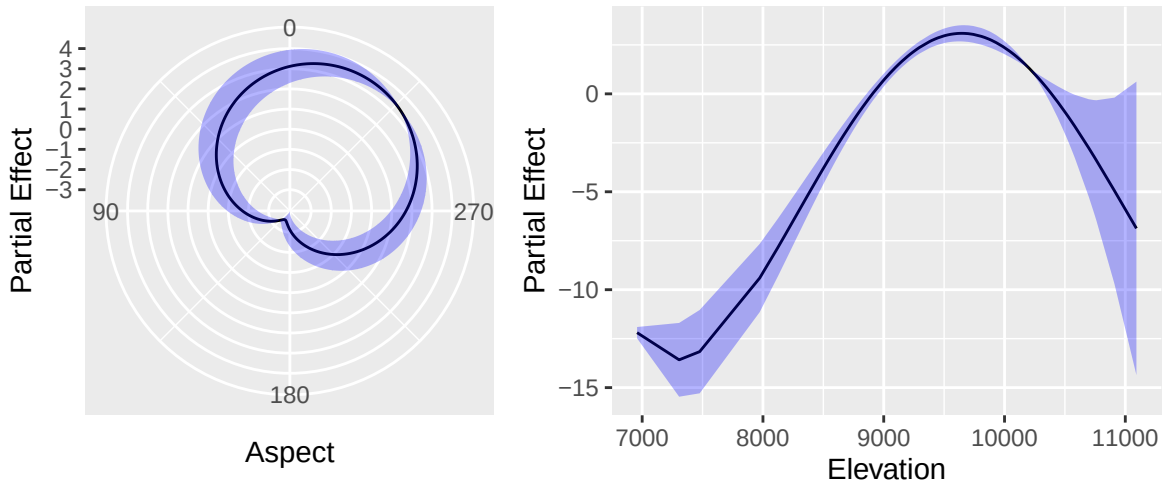

Figure 4: Partial effects plots

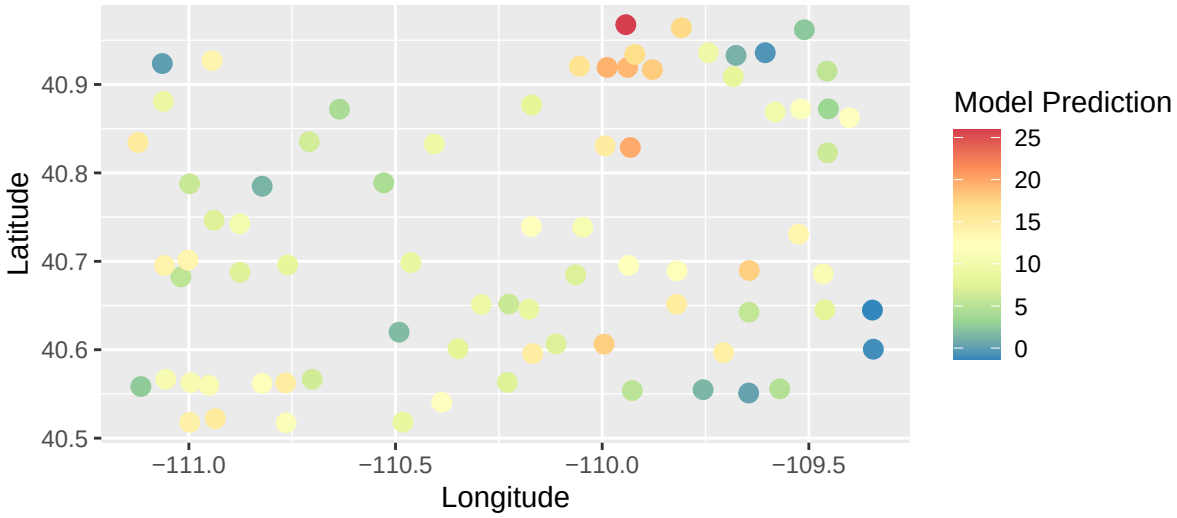

Figure 5: Model Predictions

Section 5: Conclusions

From our analysis, we found that the characteristics provided by our dataset do a good job of explaining basal area. Using our model we were able to determine that North/Northeast facing plots that are flat and between 9,500 ft. and 10,000 ft. elevation are most conducive to lodgepole pine basal area. We were also able to predict for locations that FIA was not able to measure.

One shortcoming of our model is the lack of interpretability of our orthogonal polynomial. Another is that our θ parameter confidence interval includes 0. Future analyses could explore other means of capturing the curvature evident in the relationship between lodgepole pine basal area and elevation that are more interpretable than orthogonal polynomials. There is also more modeling that could be done on the variance structure in particular. There appears ot be a variance trend in slope in addition to the non-linear variance change with elevation. This is another area future analyses could explore.