Lodgepole Pines: Nonlinear Analysis on Correlated Data

Cason Wight

March 11, 2020

Abstract

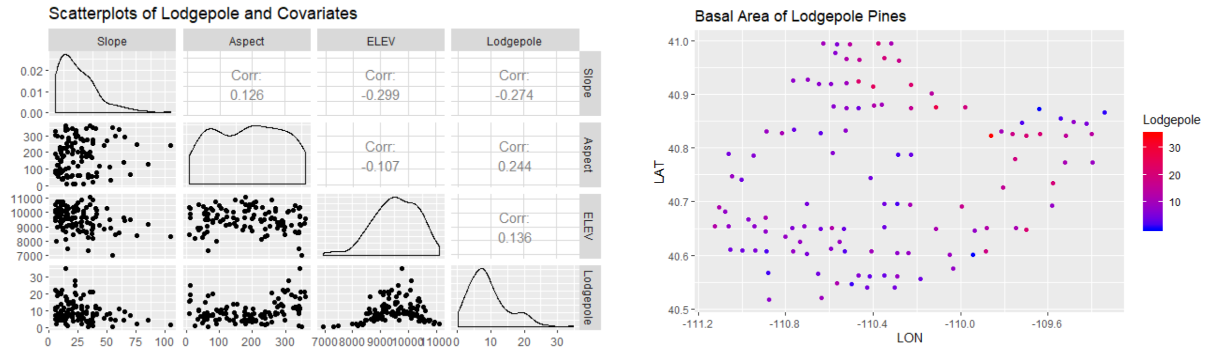
The Lodgepole pine is a versatile tree that grows in the Uinta National Forest in Northern Utah. A variety of factors have caused concern about their growth in recent years. Data collected on Lodgepole Pine basal area (in $ft^2/acre$), along with information on slope, aspect, and elevation at different locations throughout the area can reveal the effects of environmental conditions on the tree's growth. Larger slopes have a negative effect on growth, and both elevation and aspect have nonlinear effects. Predictions on unknown areas are calculated using a spatially correlated model with nonlinear effects. This model shows that the optimal condition for Lodgepole Pines is a small slope (close to 0), with a Northwestern face, at an elevation of around $9,600ft$.

1 Background

The Uinta national forest is East of Salt Lake City, just South of Wyoming. This forest is known for great camping, fishing, mountain-biking, skiing, and many other outdoor activities. This area is home to Utah's tallest mountain, King's peak. Many species, including elk, deer, goats, bears, and otters live in this area.

One vital part of the ecosystem is the pine beetles. In past years, these beetles have been essential because they make homes in dead trees and eat at them, making way for new life. Typically, the harsh winters have maintained the beetle population to reasonable levels. Unfortunately, the mild winters in recent years have caused a rampant beetle population that now feeds on both live and dead pines.

This report explores how one particular pine, the Lodgepole, is affected by particular environmental factors. The Lodgepole pine tree is versatile; surprisingly, the tree's cones actually release seeds when exposed to heat, which makes it a great tree for revitalizing a forest after a fire. Data collected on these trees reports the basal area ($ft^2/acre$), along with covariate information for coordinates (latitude and longitude), elevation, aspect (direction of face), and slope. Scatter plots on some of these covariates with the Lodgepole pine basal area, along with a plot showing the basal area records by location are shown in Figure 1.



(a) Lodgepole Basal Area and Covariates

(b) Map of Lodgepole Pine Basal Area

Figure 1: Exploratory look at Lodgepole Data

A full exploration of the effects of these covariates on Lodgepole pine basal area must account for two main data issues. First, these data have spatial correlation. One way to show this correlation is a semi-variogram, which is shown for this example in Figure 2. A semi-variogram plots the spatial dependence between points at various distances, by using a model that does not incorporate spatial correlation and comparing residuals at different distances apart.

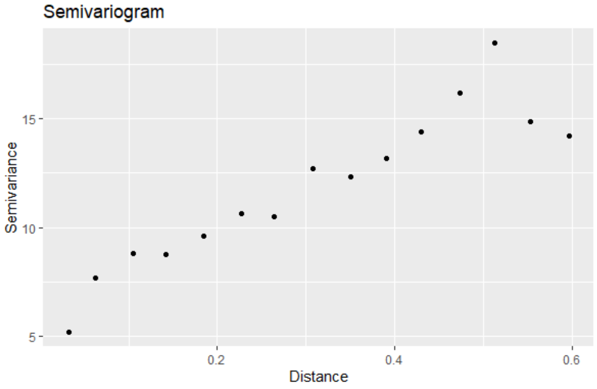
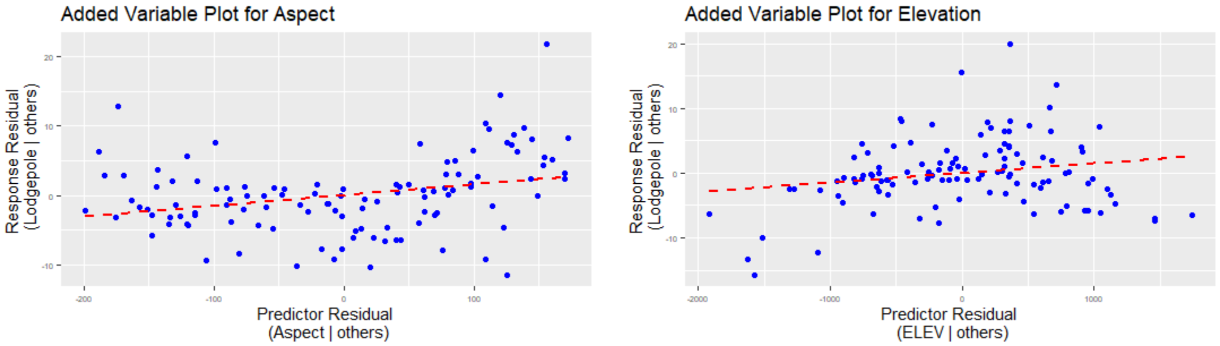


Figure 2: Spatial Dependence between Points based on Distance

The plot in Figure 2 shows lower semivariance with smaller distances, which means that measurements of basal area that are closer together have a higher correlation. The potential problem here is that without accounting for spatial correlation, the confidence and prediction intervals involving this model would be incorrect. This problem also causes invalid hypothesis test conclusions. Points for which we have a lot of information (lots of nearby points), the confidence will be too low. Points that have no nearby points will have confidence that is too high. Thus a final model must incorporate some form of spatial correlation.

Another problem with this data is that a few of the covariates have a nonlinear effect on basal area. Added-variable plots (using a simple linear model with effects of all covariates) reveal that aspect and elevation have nonlinear effects. These added-variable plots are shown in Figure 3. These plots display an increasing effect of Aspect and a decreasing effect of elevation. This means that as elevation increases, the effect of elevation on basal area decreases. If this problem is not included in a model, than the model may estimate the opposite effect of aspect and/or elevation on basal area than what it should be.



(a) Nonlinear Effect of Aspect

(b) Nonlinear Effect of Elevation

Figure 3: Added Variable Plots of Aspect and Elevation

A basis function on these two covariates can help solve this problem. Unfortunately, this report will be making predictions on some points that have elevation outside of the range of the elevation for the data for which we have basal areas. For example, the model is built off of data with a maximum elevation of $11,092$. This report will predict on points that have a maximum elevation of $11,151$ (beyond the range of the training data).

The problem here is that if a model uses a polynomial basis function expansion on elevation for prediction, the effect of elevation may diverge to a nonsensical or inaccurate amount beyond the range of the training data. A natural spline basis function expansion on this covariate will account for the nonlinear effect, but apply a linear effect beyond the bounds of the training data, thus making for more reasonable predictions.

Aspect is a measurement of which direction a location is facing. A measurement of 90 means that at that point, it is facing North. A simple cosine/sine basis function expansion will convert this value into directional values. Figure 4 shows the basal areas based on the aspect. If this type of basis function expansion is not applied, then an aspect of 1 will be measured as extremely different than 359, even though in reality those two aspects are extremely similar. If this problem was not accounted for, the difference in effects of two aspects may be much larger than it should be.

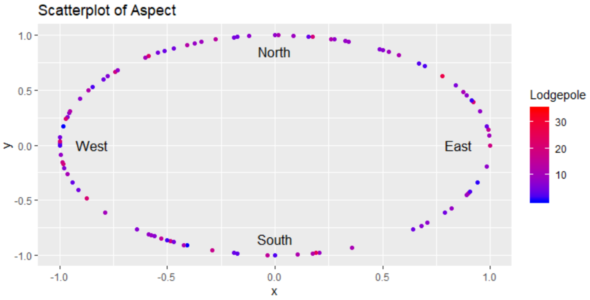


Figure 4: Lodgepole basal area by Aspect

The goals of this report are the following:

1. Define a model that accounts for spatial correlation and nonlinear effects
2. Verify that the assumptions of the model and assess fit and prediction ability
3. Explore how Lodgepole basal area is affected by its environment
4. Identify what conditions are conducive to Lodgepole pine growth
5. Predict the basal area for Lodgepole pines for areas with unknown basal areas

2 Introduction of Methods

The method that will account for both the spatial correlation and the nonlinear effects is a spatial Gaussian process with basis function expansions for aspect and elevation. The two pieces of this method (spatial Gaussian process and basis function expansions) will be explained, along with their strengths and weaknesses. The assumptions of this method are also defined.

A natural spline is a basis function expansion, meaning it expands an **X** matrix to include new columns that expand the original columns. A natural spline fits a third-degree polynomial on various segments of the original column. The segmentation points are known as *knots*. Natural splines impose the constraint the third-degree polynomial has a continuous first and second derivative at the knots. This constraint leads to continuous changes in slope and continuous changes in convexity, essentially ”smoothing out” the function. Natural splines will be used to get a basis function expansion on elevation.

A great advantage of natural splines is that outside of the boundaries of the knots, it fits a linear trend, which limits the divergence that is likely when fitting a simple polynomial basis function expansion. Natural splines allow for a flexible polynomial fit within the bounds and a less extreme, linear fit outside of the bounds of the knots. One of the downsides to natural splines is that the number of knots (and where to place the knots) is difficult. Cross validation is used to find the optimal number of knots that will be equally spaced on the data. The actual function for the basis function expansion is complicated; this leaves the model terms without any reasonable interpretation. Plots are the only way to show the effect of a covariate on the response.

The basis function expansion that will be used for aspect is a cosine/sine function. The goal is to get a continuous, circular scale for aspect, that measures directions properly. The scale of 0 to 360 is changed to radians, by multiplying aspect by $\frac{2\pi}{360}$. Adding $\pi/2$ will set the index so that an aspect of 0 is taken as North in a plot. Taking the cosine of this value gives a West/East measurement and the sine of this value gives a North/South measurement. Therefore aspects that are similar to each other are now forced to have similar effects.

A Gaussian process is a method of fitting a set of data as a single observation of a multivariate normal distribution, with a specified variance-covariance matrix. The model is defined by the following:

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

σ^2 is the variance of decorrelated residuals and ω is the "nugget", or level of sampling variability. **I** is simply the identity matrix. In this example, the **R** matrix is defined as a correlation matrix between the residuals of the points in space, defined by the observations of the data.

$$\mathbf{R} = \begin{bmatrix} 1 & \text{Corr}(\epsilon_1,\epsilon_2) & \text{Corr}(\epsilon_1,\epsilon_3) & \dots & \text{Corr}(\epsilon_1,\epsilon_n) \\ \text{Corr}(\epsilon_1,\epsilon_2) & 1 & \text{Corr}(\epsilon_2,\epsilon_3) & \dots & \text{Corr}(\epsilon_2,\epsilon_n) \\ \text{Corr}(\epsilon_1,\epsilon_3) & \text{Corr}(\epsilon_2,\epsilon_3) & 1 & \dots & \text{Corr}(\epsilon_3,\epsilon_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Corr}(\epsilon_1,\epsilon_n) & \text{Corr}(\epsilon_2,\epsilon_n) & \text{Corr}(\epsilon_3,\epsilon_n) & \dots & 1 \end{bmatrix}$$

The correlation between the residuals is defined as a function of the euclidean distance between two observations. The ϕ term (the range) manages the correlation strength based on the distance and is chosen through maximum likelihood estimation.

$$\text{Corr}(\epsilon_i,\epsilon_j) = \exp\left\{-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{\phi}\right\} = \exp\left\{-\frac{\sqrt{(\text{Lon}_i - \text{Lon}_j)^2 + (\text{Lat}_i - \text{Lat}_j)^2}}{\phi}\right\}$$

The $\boldsymbol{\beta}$ matrix is a $k \times 1$ matrix of the linear effects of the columns of the $n \times k$ **X** matrix, which is defined by the following:

$$\mathbf{X} = \begin{bmatrix} 1 & \text{Slope}_1 & \cos\left(\frac{2\pi}{360}\text{Aspect}_1 + \pi/2\right) & \sin\left(\frac{2\pi}{360}\text{Aspect}_1 + \pi/2\right) & \mathbf{f}(\text{Elevation}_1) \\ 1 & \text{Slope}_2 & \cos\left(\frac{2\pi}{360}\text{Aspect}_2 + \pi/2\right) & \sin\left(\frac{2\pi}{360}\text{Aspect}_2 + \pi/2\right) & \mathbf{f}(\text{Elevation}_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

where **f** is the multi-column natural spline function. This function is too complex to show above.

The strength of using basis function expansions is that it allows for nonlinear effects, while using a linear model. For elevation, the natural spline is particularly beneficial because this report predicts on areas with elevation outside the range of the elevation of the training data. Natural splines fit a linear trend beyond the range of training data, so there will not be divergent predictions outside of that range. Within the range, a smoothed piecewise polynomial is fit. The downside to this method is that the effects (in the $\boldsymbol{\beta}$ matrix) are not interpretable.

The cosine/sine basis function expansion on aspect helps with interpretation. Changing the range of 0 to 360 to radians with range 0 to 2π allows the cosine and sine functions to measure the level of North/South and West/East aspect. These values are more interpretable and allow for a continuous variable instead of a ranged aspect variable. This basis function expansion is more difficult to plot, and can still be hard to understand.

The spatial Gaussian process models spatially correlated values, which is what is needed in this approach. Well outside the range of the training data, predictions rely more on $\mathbf{X}\boldsymbol{\beta}$ and near training data, predictions rely more on the known data. Unfortunately, the weakness is that this process takes more time than a simple linear regression with an independent response.

This model still assumes a linear relationship in $\boldsymbol{\beta}$ on the response, even after the basis function expansions. Independence is not assumed in this model, but the decorrelated residuals should still be normally distributed and have equal variance.

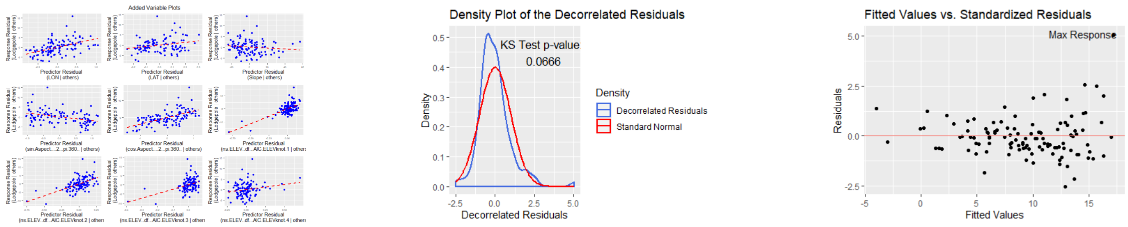
3 Justification and Performance Evaluation

The exact model included had 4 knots for elevation and excluded none of the variables. The knots were selected based on root mean squared error (RMSE) through cross validation. The best value for ϕ and the best nugget value is found through maximum likelihood.

A likelihood ratio test determined that elevation was has a significant nonlinear effect on Lodgepole basal area (p-value: $< .001$). Aspect was also found to have a nonlinear effect (p-value: $< .001$). A likelihood ratio test also confirmed the importance of including a spatial correlation term (p-value: $< .001$).

As stated in the previous section, the three assumptions that have to be met are linearity of $\boldsymbol{\beta}$ effects on the response, normality of decorrelated residuals, and equal variance of decorrelated residuals.

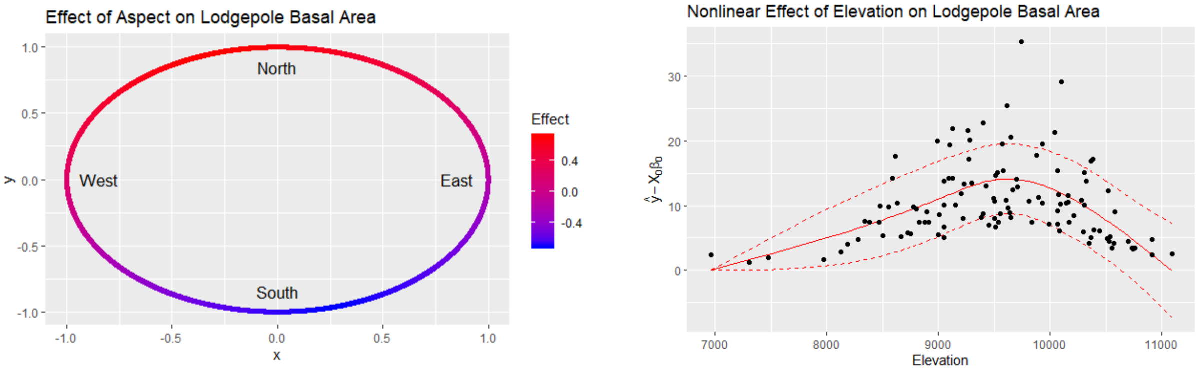
The linearity can be checked through an added variable plot. Any curvature would indicate nonlinearity in the $\boldsymbol{\beta}$. Normality can be checked by a density plot of the decorrelated residuals, compared to a standard normal density. The equal variance assumption can be checked by a scatterplot of residuals against the fitted values. There shouldn't be an increasing or decreasing spread.



(a) Linearity (b) Normality (c) Equal Variance

Figure 5: Checking the Assumptions

Figure 5 contains all three of these plots. The added variable plots have no concerning curvature. The density plot looks roughly normal. A KS test yields a p-value of .0666 of the decorrelated residuals being significantly different from a Normal distribution. The fitted values vs residuals are slightly conical, but not enough for major concern. The outlier shown in the top right corner is observation 35, which is the observation with the highest Lodgepole basal area. There is no extreme or concerning evidence against any of the three assumptions.



(a) Effect of Aspect on Lodgepole basal area (b) Effect of Elevation on Lodgepole basal area

Figure 6: Nonlinear Effects

The cross-validated RMSE of this model is 2.8035. Compared to the standard deviation of the original areas of 6.2391, this is small prediction error. Instead of an R^2 value, this report includes a substitute $R^2_{pseudo} = \text{cor}(\hat{y}, y)^2 = 0.5951$. This means that there is a $\sqrt{R^2_{pseudo}} = \sqrt{.5951} = .7714$ correlation between fitted and actual values using this model. The maximum R^2_{pseudo} is 1.000 and the minimum is 0.0000. Thus a R^2_{pseudo} of 0.5951 is relatively good fit.

4 Results

The two main questions of this report are 1) what are the optimal conditions for Lodgepole pine growth and 2) what are the predicted basal areas for the locations for which they are unknown.

Based on the model, for a 1 degree increase in slope, the Lodgepole pine basal area decreases by 0.0391 (95% CI: 0.0039 to 0.0742). A bigger slope can lead to soil erosion and shallow roots. The two covariates with nonlinear effects are interpreted graphically. Figure 6 shows these effects. On the left, the plot shows the effect of different aspects on Lodgepole basal area, ranging from -0.7034 to 0.7034. Because the sun sets in the West, a Westward tree likely gets more sun. On the right, the plot shows the effect of elevation on Lodgepole basal area. The elevation is a balancing act because fewer species can survive at higher elevations, so some elevations may be conducive for the Lodgepole and many other species, while others just for the Lodgepole. At a certain elevation the Lodgepole pine also struggles to grow.

Thus, the best conditions for Lodgepole pine growth is a slope as small as possible (0 degrees), the mountain facing Northwest, and the elevation about $9,600ft$. Because areas are spatially correlated, a location with good growth will likely have good growth in nearby locations.

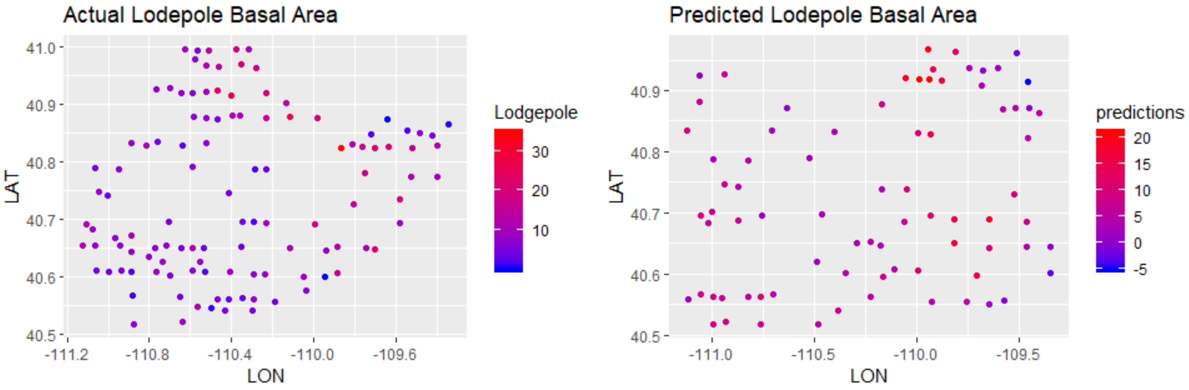


Figure 7: Predicted Lodgepole basal area

Figure 7 shows the predictions on the specified locations with no numbers for Lodgepole basal area. The predictions are compared side-by-side with the actual given values, for comparison. The 95% prediction intervals are shown in Figure 8.

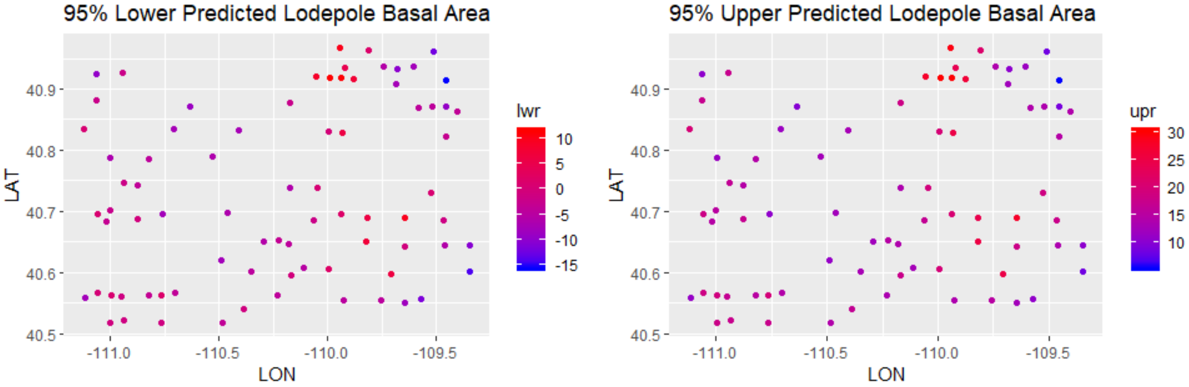


Figure 8: 95% Prediction Intervals on Lodgepole basal area

5 Conclusion

The goals of this report started with defining a model that accounts for spatial correlation and nonlinear effects. The spatially correlated Gaussian process with basis function expansions for elevation and aspect did this. The assumptions of this model were linearity of $\boldsymbol{\beta}$ on the response, normality of decorrelated residuals, and equal variance of the decorrelated residuals. The model had an RMSE of 2.8035, which is low when compared to the standard deviation of 6.2391. The R^2_{pseudo} was 0.5951.

Larger slope has a negative effect on Lodgepole basal area. Elevation has a positive effect on basal area at lower elevations and a negative effect at higher elevations. Northwestern aspect has a positive impact on basal area, while Southeastern has a negative effect. The environmental factors most conducive to Lodgepole pine growth is a flat area, preferably facing the Northwest, with an elevation of about $9,600ft$. Predictions for the unknown areas are shown in Figure 7, with 95% prediction intervals shown in Figure 8.

There are a few shortcomings to the current method. One shortcoming is that other forms of spatial correlation besides Gaussian with equal variance in decorrelated residuals was not tested. The density plot of the residuals and the fitted values vs. residuals plot revealed that checking some other form of model is warranted, although the current approach is not entirely unrealistic.

Another shortcoming of this analysis is that it encapsulates little about the environmental factors. Even though the R^2_{pseudo} isn't bad, there are likely many other factors that could play into how well Lodgepole pines grow. Covariates concerning wildlife growth, water flow, air quality, and others could give a more comprehensive look into Lodgepole growth. This could be a next step for the analysis. Other next steps for the analysis would be evaluating other non-Gaussian forms of correlation or how these responses change over time.