Analysis of Cartographic Features and Forest Covertype

Mikkel Bjornson

10/9/2021

Introduction

Dominant species in tree cover can affect the structure of an ecosystem. The topographic features of that environment may effect which species dominate. Initial analysis focuses on the effects of elevation and other topographic features on the presence of spruce and fir trees. By quantifying these effects, insights into the niche of these trees can be better understood. The effects of elevation on tree species are of particular interest. In addition, Mapping tree species over large areas can be difficult and costly. Associations between tree species and topography may help estimate the forest cover. The topographic features are explored in relation to predicting forest cover type.

Data Description

The analysis uses the cover type data (Bache and Lichman 2013). Four wilderness areas within the Roosevelt National Forest are sampled using remote mapping techniques. Cartographic features are obtained for thirty by thirty meter plots. Cover type is obtained from the USFS resource information center. The data contains twelve explanatory variables and one response. Variables include:

Variable Name	Description
Elevation	Meters above sea level that the plot is located
Aspect	Degress from north that the slope of the plot faces
Slope	Steepness of a hill the plot is on, 0 indicating flat
Horz_dist_hydro	Horizontal distance in meters to the nearest body of water
Vert_dist_hydro	Vertical distance in meters to the nearest body of water, negative values indicating water at lower elevations
horz+dist_road	Horizontal distance in meters to the nearest road
Hillshade9	Calculated illumination of the plot at 9am on the summer solistice, based of topographical features, 0 indicating shaded plot
Hillshade12	Calculated illumination of the plot at noon on the summer solistice, based of topographical features, 0 indicating shaded plot
Hillshade3	Calculated illumination of the plot at 3pm on the summer solistice, based of topographical features, 0 indicating shaded plot
Horz_dist_fire	Meters to the nearest forest fire starting point
wild_area	Indicates which of the four wilderness areas the sample comes from

Variable Name	Description
soil_type	Indicates which of the 40 different soil types is present in the plot
cover_type	Indicates the dominant tree species on the plot

Analysis begins with an inferential study of the effects of elevation on spruce/fir cover when controlling for other cartographic variables. There is some concern about lack of independence between observations. The large number of observations additionally provide very large degrees of freedom. To help alleviate both problems, inference is conducted on a random sample of 5000 observations. The cover type variable is reduced to a binary response. Aspect is reduced to a factor variable indicating cardinal or primary intercardinal direction.

Initial exploration, revealed apparent differences in elevation for each the seven cover types. There is also apparent differences between soil and cover type. However, there are no recorded instances of spruce cover in Montane alluvial soil (type elu51). It is unknown whether spruce cannot grow in this soil type or it is just not observed. The Cache wild area also has no plots with spruce trees. The perfect separation prevents the use of maximum likelihood methods, and thus was thus excluded from the logistical regression analyses. Examination of the correlation matrix and scatter plot matrix finds some multicolinearity between Hillshade12 and Slope (r = -0.5199), Vert_dist_hydro and Horz_dist_hydro (r= 0.5942), and Hillshade_3 and Hillshade_9 (r= -0.7791). Hillshade12, Horz_dist_hydro, and Hillshade_9 were dropped as they had larger correlation coefficients with other explanatory variables.

Logistic Regression methods are selected for the inferential analysis (Kutner et al 2004, James et al 2021). A full model is fit including all remaining explanatory variables and interactions between elevation, slope, and aspect of the hill. Insignificant variables and interactions are dropped from the model. The drop in deviance test is used to determine appropriateness of dropping variables. The final model is the remaining reduced model.

Further analysis focuses on the predictive power of cartographic features. Both a multiple logistic regression models using a one vs. rest method (Kutner et al 2004, James et al 2021), and a multiclass random forest (James et al 2021) are explored. The covertype data is split into training(60%), validation(20%), and test(20%). The multiple logistic regression models split the cover type variable into 7 binary response variables. Logistic regression models are built for each cover type, using a classification tree to select explanatory variables. Predictions are determined by the model with the largest response value. The several random forest models are built with varying numbers of trees and features to split. The model with the lowest validation error is selected.

Statistical Modeling

Inferential model

Three models are fit to the data set for the purpose of estimating the effect of elevation on Spruce/Fir tree cover. An initial rich model is fit using all available variables.

$$\begin{aligned} logit(spruce/fir) = \beta_0 + \beta_1(Elevation) + \beta_2(Aspect) + \beta_3(Slope) + \beta_4(Vert\_dist\_hydro) + \dots \\ \beta_5(horz\_dist\_road) + \beta_6(Hillshade3) + \beta_7(Horz\_dist\_fire) + \beta_8(Elevation * Aspect) + \dots \\ \beta_9(Elevation * Slope) + \beta_{10}(Aspect * Slope) + \beta_{11}(Elevation * Aspect * Slope) \end{aligned}$$

Two reduced models are also evaluated:
- reduced model 1:

$$\begin{aligned} logit(spruce/fir) = \beta_0 + \beta_1(Elevation) + \beta_2(Aspect) + \beta_3(Slope) + \beta_4(Vert\_dist\_hydro) + \dots \\ \beta_5(horz\_dist\_road) + \beta_6(Hillshade3) + \beta_7(Horz\_dist\_fire) + \beta_8(Aspect * Slope) \end{aligned}$$

- reduced model 2:

$$\begin{aligned}logit(spruce/fir) = \beta_0 + \beta_1(Elevation) + \beta_2(Aspect) + \beta_3(Slope) + \beta_4(Vert\_dist\_hydro) + \dots \\ \beta_5(horz\_dist\_road) + \beta_6(Hillshade3) + \beta_7(Horz\_dist\_fire) + \beta_8(Aspect * Slope)\end{aligned}$$

Predictive model

The first predictive model is a multiclass logistic predictive model fit the following logistic models:

$$logit(spruce/fir) = \beta_0 + \beta_1(Elevation)$$

$$logit(Lodge\ pole\ pine) = \beta_0 + \beta_1(Elevation)$$

$$\begin{aligned}logit(spruce/fir) = \beta_0 + \beta_1(Elevation) + \beta_2(Vert\_dist\_hydro) + \beta_3(horz\_dist\_road + \dots \\ \beta_4(Hillshade9) + \beta_5(Hillshade3)\end{aligned}$$

$$\begin{aligned}logit(cottonwood) = \beta_0 + \beta_1(Elevation) + \beta_2(Horz\_dist\_hydro) + \beta_3(horz\_dist\_road) + \dots \\ \beta_4(Hillshade9) + \beta_5(Horz\_dist\_fire) + \epsilon\end{aligned}$$

$$logit(Douglas\ fir) = \beta_0 + \beta_1(Elevation)$$

$$logit(Krummholz) = \beta_0 + \beta_1(Elevation)$$

$$logit(PonderosaPine) = \beta_0 + \beta_1(Elevation) + \beta_2(Hillshae12)$$

Using a one vs. rest strategy, the model producing the largest response indicates the predicted value.

The second predictive model is a Random Forest model with 200 trees.

Results

Inferential study

The second reduced model is chosen over both the full and first reduced models using the drop in deviance test (Full model vs reduced model 2: p-value<0.1669). The table below indicates Elevation and Vertical distance to water have significant responses. There is weak evidence of significant response to slope and interaction between slope and aspect. Isolating elevation, there is evidence to suggest that elevation is associated with the presence of spruce/fir as the dominant tree cover (p-value < 0.00001). With 95% confidence, the logistic model estimates the presence of spruce/fir cover increases by between 74.6% and 88.59% for every 100 meter increase in elevation. There is also an estimated decrease in the odds of spruce and firs trees by between 24.3% and 33.53% per 50m increase in vertical distance to water with 95% confidence.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-18.36732	0.63550	-28.90216	0.00000
Elevation	0.00595	0.00020	30.27681	0.00000
AspectNE	0.33366	0.25777	1.29441	0.19552
AspectE	0.19709	0.26620	0.74038	0.45907

	Estimate	Std. Error	z value	Pr(>|z|)
AspectSE	0.19566	0.28760	0.68032	0.49630
AspectS	-0.31468	0.32402	-0.97115	0.33147
AspectSW	-0.23493	0.32328	-0.72673	0.46739
AspectW	-0.24984	0.30945	-0.80737	0.41945
AspectNW	-0.28177	0.28172	-1.00019	0.31722
Slope	0.02552	0.01312	1.94542	0.05172
Vert_dist_hydro	-0.00686	0.00066	-10.34489	0.00000
AspectNE:Slope	-0.01413	0.01790	-0.78948	0.42983
AspectE:Slope	-0.02469	0.01824	-1.35362	0.17586
AspectSE:Slope	-0.04786	0.01997	-2.39613	0.01657
AspectS:Slope	-0.03034	0.02248	-1.34985	0.17706
AspectSW:Slope	-0.02919	0.02114	-1.38082	0.16734
AspectW:Slope	-0.00089	0.02023	-0.04418	0.96476
AspectNW:Slope	0.03680	0.01891	1.94579	0.05168

Predictive Model

The multiclass logistic regression model achieved a test error rate of 0.3526. Examining the per class accuracy rates in the table below, the model accurately estimated Lodge pole pine (accuracy$\approx$ 81%) and Spruce/fir (accuracy$\approx$ 64%). It suffered from very poor accuracy on the less common cover types.

The Random Forest model appears to fit the data better with a test error rate of 0.0344. The per class accuracy rate is above 90% for most cover types. Cottonwood/willow (accuracy$\approx$ 87%) and Aspen (accuracy$\approx$ 86%) are the least accurately predicted.

Cover	Logistic	Random_Forest
Spruce/Fir	0.6387	0.9622
Lodgepole Pine	0.8105	0.9765
Ponderosa Pine	0.2885	0.9673
Krummholz	0.0327	0.9578
Cottonwood/Willow	0.0000	0.8482
Aspen	0.0000	0.8449
Douglas-fir	0.0000	0.9182

Conclusion

Elevation and other cartographic features appear to be associated with the dominant tree species in forests of Roosevelt National Forest. There is an estimated increase in the chance of finding spruce/fir between 74.6% and 88.59% per 100 meter increase in elevation. This suggests that spruce/fir forests perhaps favor higher elevations. The negative association with vertical distance to water suggests spruce and fir will only prefer the higher elevations as long as they are not too far from water. Although the other variables were not found to be significant, the drop in deviance test suggest they are still important factors. The possible lack independence between the observations and the removal of variables from analysis leave some room for skepticism while interpreting these results.

When attempting to predict forest cover, the logistic model failed to capture the complexity of the system with an accuracy rate of about 65%. The Random Forest model managed about a 96% success rate. The ability of random forest models to manage non-monotonic changes is likely a factor in the higher success rates. Further model building using xgboost and neural networks could possibly increase the predictive accuracy even further.

Citations

Blackard, Jock A. 1998. “Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types.” Ph.D. dissertation. Department of Forest Sciences. Colorado State University. Fort Collins, Colorado. 165 pages.

Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/covertype. Irvine, CA: University of California, School of Information and Computer Science

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied Linear Regression Models (fourth). McGraw-Hill Irwin.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning with applications in R. Springer.