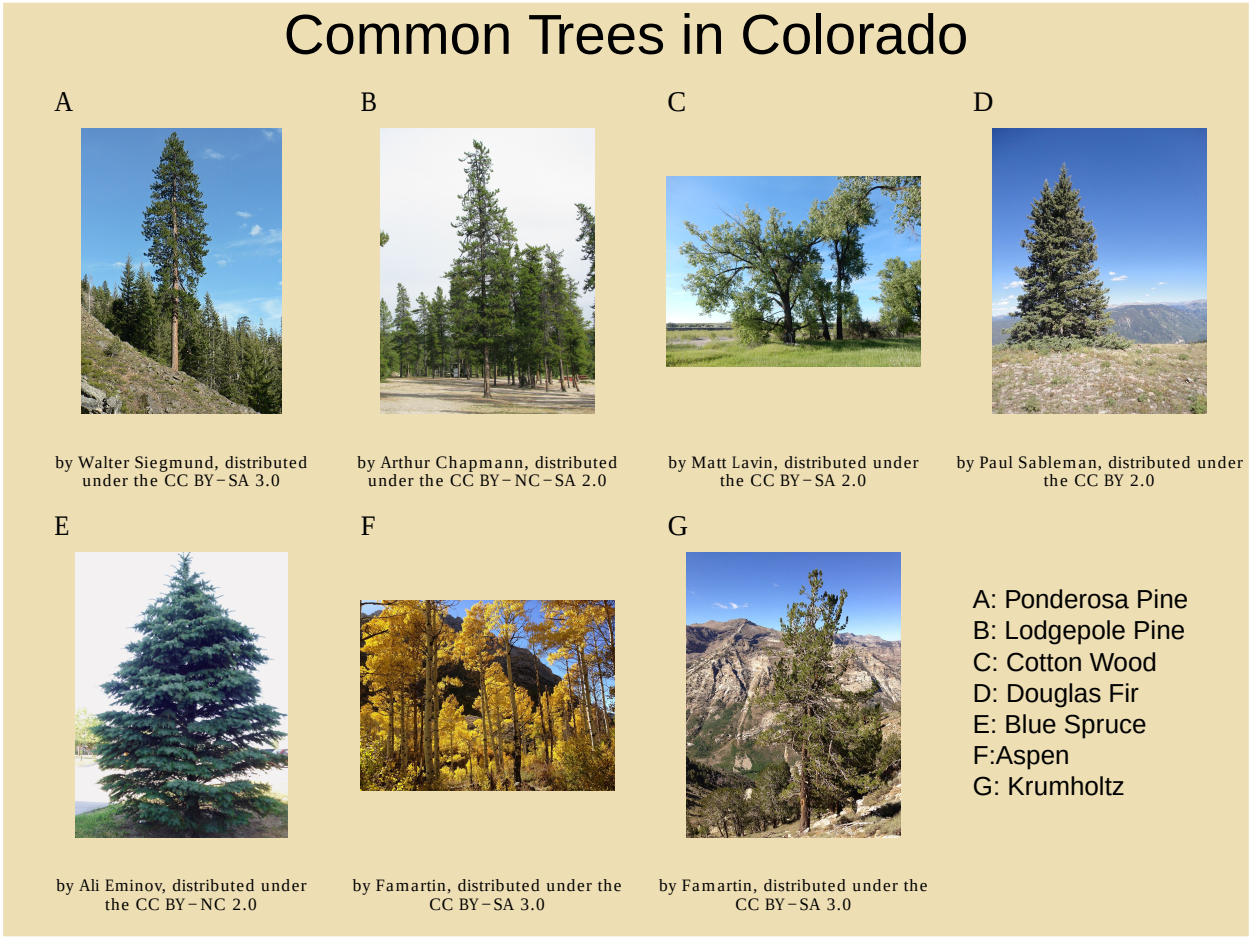
Guessing the Types of Trees by Looking at Maps

Mikkel Bjornson

10/9/2021

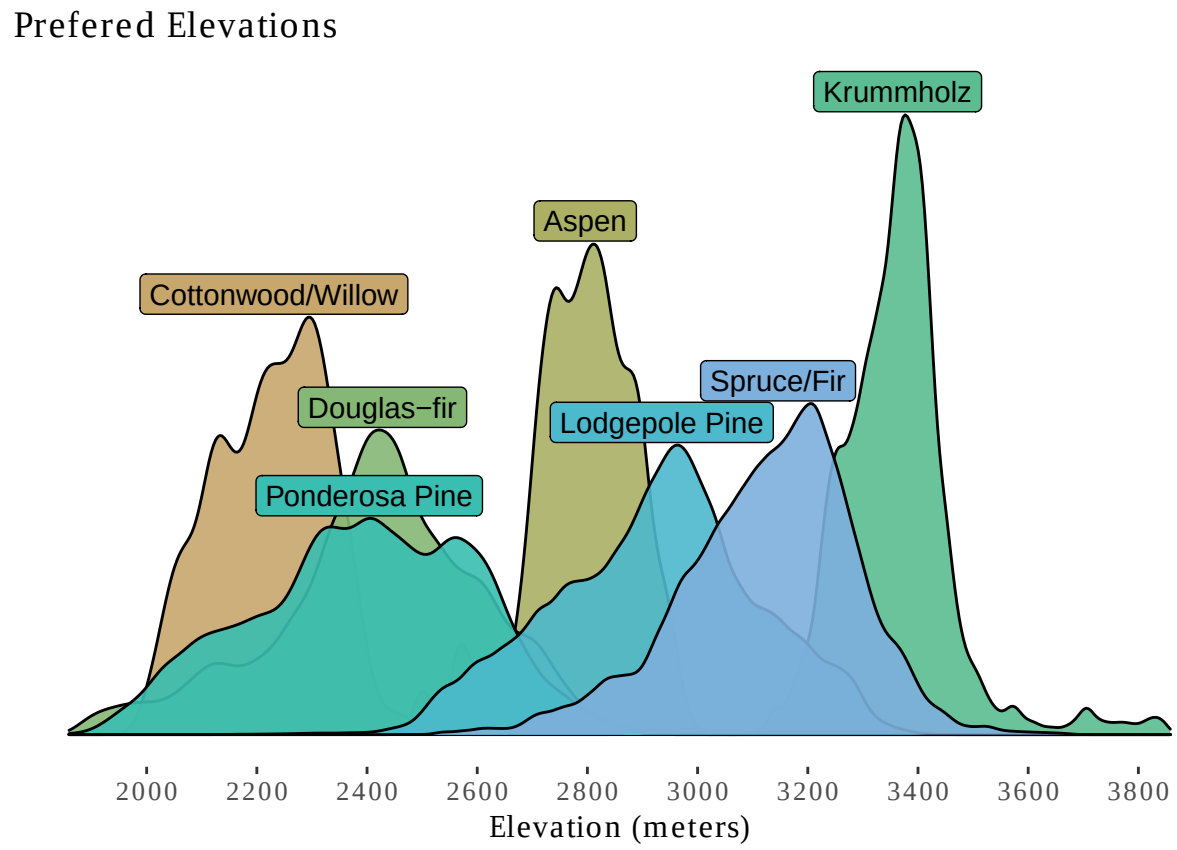
Driving across Colorado, we experience environments from Grasslands and semi-desert shrub land to montane forest and alpine tundra. The state has a diversity of habitats each with their own plants and animals. We focus on the trees within these habitats. The goal is to know where each species of tree is located. However, Imagine the difficulty of mapping each tree individually. Even visiting all the trees in a national park would be a time consuming ordeal. Add to that the time and expense of recording the species and location and it becomes unpractical. Knowing this information exactly, could extend our knowledge of the ecosystem and effects of climate change. It would also be useful in land management, but not at this expense. Since we cannot get this kind of exact data, the next best thing is a good estimate.



To make estimates, we start with a publicly available cover type data set (Bache & Lichman 2013). It contains records for 30 meter by 30 meter plots. Many physical characteristics such as the elevation of the plot, slope of the land, and distance to water are easily found on maps. A total of twelve different

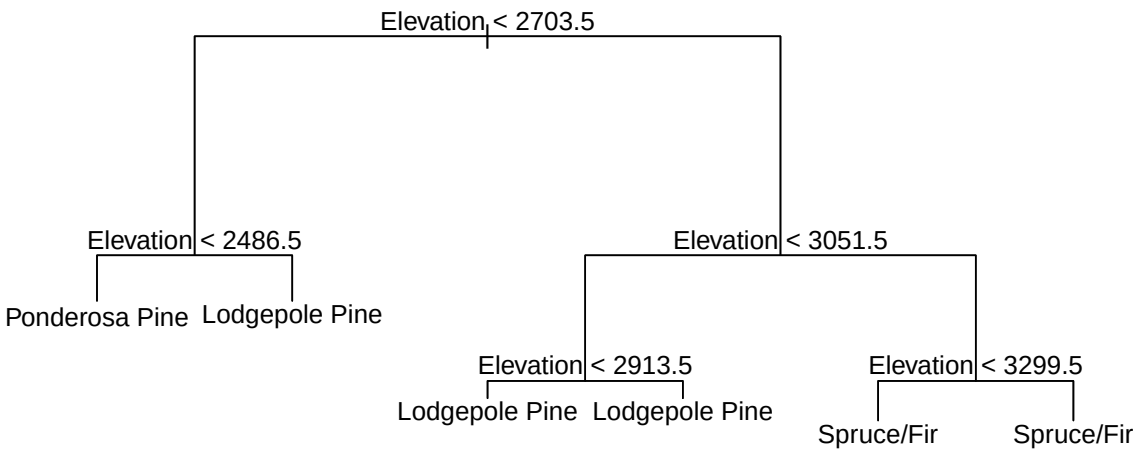
characteristics were recorded. These measurements were made for over 580,000 plots. This would cover an area slightly smaller than Redwood National Park. In addition, the dominant tree species was obtained from USGS and USFS data. A total of seven different tree species were recorded.

Comparing the information from the maps with the data about tree species, all 12 characteristics recorded played some role in tree type. Elevation stood out as one of the most important. Each species appears to have a range of preferred elevations. This strong relationship is good news. It provides a powerful tool when attempting to estimate the tree type. For example, if a plot of land has an elevation of about 3200 meters, we can look at the graphs below and see that spruce/fir are the most common type and therefore most likely. However, we can do even better at these predictions by incorporating more sophisticated methods.



Two different methods were attempted to make predictions. The first method is multi-class logistic regression. This model predicts the odds a plot has each type of tree. We then classify the plot as containing the tree with the largest odds. Unfortunately, this model did not perform very well. It classified only about 35% of the plots accurately. This is still more accurate than randomly guessing one of the seven tree types, but not really useful. Further investigation revealed that this likely happened because logistic regression is not flexible enough. It allows the odds of a tree species to go up as elevation goes up, but it does not allow those odds to go back down once the elevation is beyond the ideal level.

The second method is Random Forest Classifier and it is much more flexible. Using a computer, a sample of the data is selected and used to make a decision tree. The computer did this a total of 200 times. Below is an example of a decision tree. The tree splits as we travel down based upon different characteristics of the plot. To make a prediction we follow the splits based upon the values for the plot until we get to a tree species. The Random Forest model uses all 200 trees, makes 200 estimates, then averages the estimates together. This allowed the estimates to be correct 96.5% of the time.



Despite the good overall estimation rate, some species are predicted better than others. Lodgepole and Ponderosa pine were both predicted with over 97% accuracy. Cottonwood was predicted at about 87% accuracy, and Aspen at only 84.7% accuracy. The large difference in per species prediction rates leaves room for more improvement. Possible paths toward better accuracy include finding and using more map features, and the application of more powerful methods. Further research will likely yield even better predictions.

Citations

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