# Utilizing FIA Data & Machine Learning to Improve Forest Health


PURDUE UNIVERSITY | The Data Mine

Ethan Nie, Fletcher Vielee, Sienna Grey, Elisabeth Dunch, Siddarth Nachannagari, Sophia Chiparus, Claire Mui


## OBJECTIVES & MOTIVATIONS

- Explore the FIA (Forest Inventory Analysis) database
- Utilize R and SQL code through Anvil (an online workspace) to understand how forest vegetation responds to restrictions placed on them (first semester we focused on data with and without forest fires, and whether they were private or not).
- Create models using the code developed from R-studio code.
- Use these models and graphs to analyze the tree species, or types within a certain region.
- Utilize Models such as KNN to develop models that account for areas with missing trees, to allow for overall analysis of the forest of a state or country.
- Use Machine Learning and KNN to answer how ecological patterns impacts overall forest vegetation patterns.
- Utilize Machine Learning models to easily analyze tree species vegetation, check for tree species relation, and determining forest type with lack of tree species data in a set region.
- Compare the models to understand which external factor ranks in terms of determining the vegetation.

## NETWORK GRAPH ANALYSIS

The previous semester our team worked to utilize network graph analysis to show connectiveness and correlation between different plots.
We specifically looked at species codes to see how they correlate to one another. For example, do certain trees commonly appear together, or apart? What these kinds of graphs essentially help show is interconnectivity. The more lines or clutter between species, the more connected they are. We then used different visualizations to help focus on the amount of connectivity.
Then we tested our code by analyzing tree species in specific areas, such as those prone to forest fires in the west coast, or those that are more prevalent in different altitudes. This was to understand how environmental conditions impact the overall species density of a plot region. A network graph can also be an effective way to visualize the biodiversity of a region.

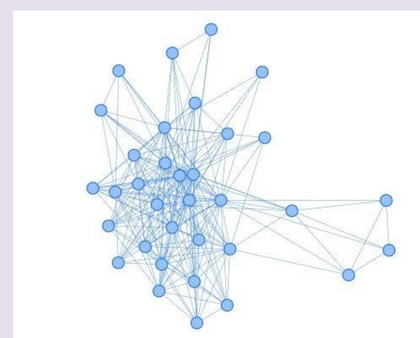

Fig. 1: An example of a network graph.

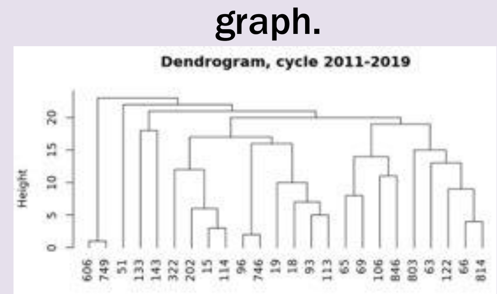


**Fig. 1.1: A dendrogram of tree species in Coconino County, Arizona.**

## WHAT IS KNN (K-NEAREST NEIGHBOR)?

KNN code was used to build a predictive model that predicts the type of forest (StableForest) in our data set. The code accomplishes this by comparing one unknown point with the k nearest points and uses this to classify the unknown point. For variable pairs that showed a high correlation in the matrix, one variable could be left out to increase accuracy. Other variables, after being log transformed, also resulted in a higher model accuracy.
Our goal was to determine which variables, helped the predicative model accomplish the highest accuracy, which should be left out, which should be transformed, and what value of K to use.

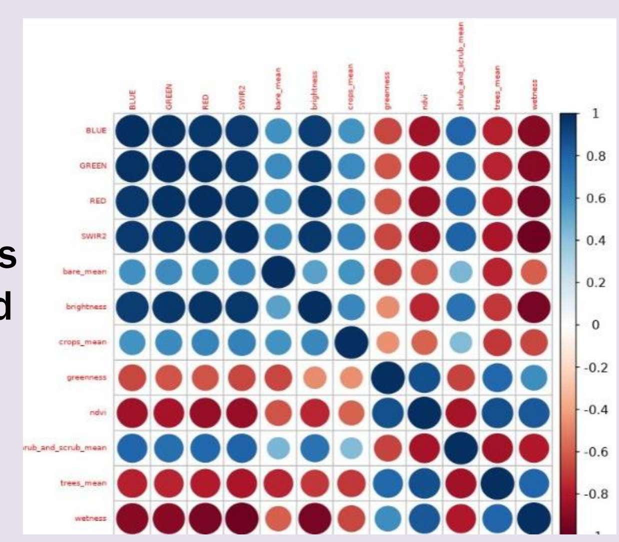

**Fig. 2: A correlation matrix of variables used in the StableForest predictor model**

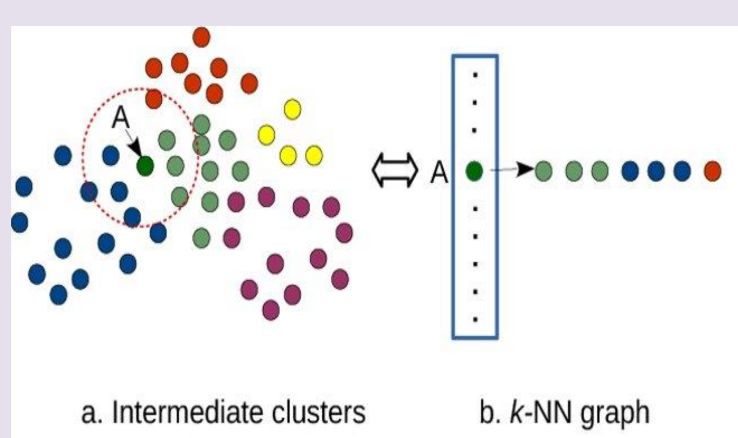


**Fig. 2.1: A graphic showing how KNN works.**

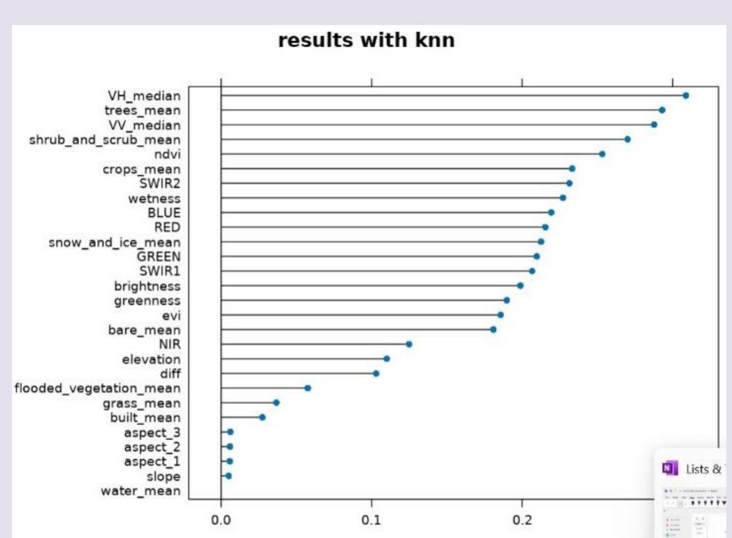


**Fig. 2.2: Shows the importance levels of influence each variable has on the model.**

## WORKING WITH FIA DATA

The USDA's Forest Inventory and Analysis (FIA) program collects and analyzes data important to assessing the condition of forests across the United States. The database consists of hundreds of tables, which contain data on hundreds of variables. Commonly used tables for this project include the tree and plot tables. Commonly used variables are tree height, diameter, and species, as well as plot data. Our team set out to parse and manipulate the data, utilizing various coding languages and applications.

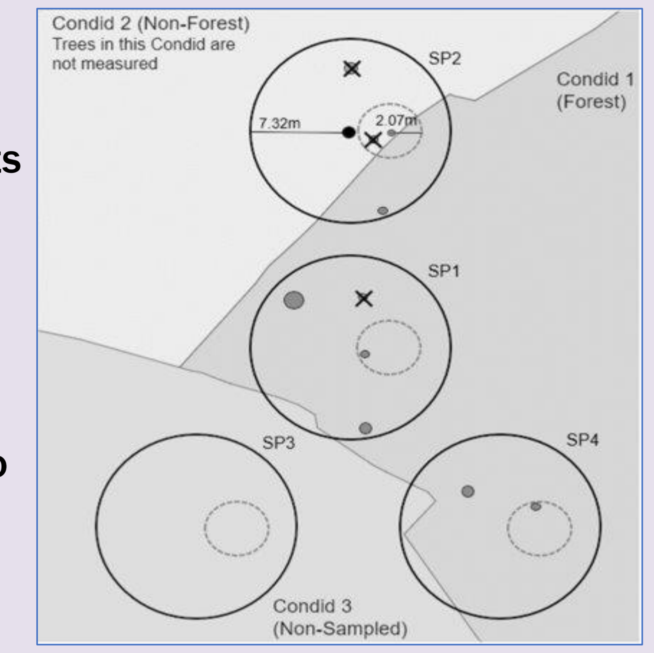


**Fig. 3: A graphic showing how the FIA database categorizes and labels certain aspects of plots.**

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| | SP1 | SP2 | COMMON_NAME_SP: | COMMON_NAME_SP: | NUMPLOTS | NUMPAIRS |
| 1 | 68 | 817 | eastern redcedar | shingle oak | 1 | 10 |
| 2 | 68 | 802 | eastern redcedar | white oak | 1 | 4 |
| 3 | 68 | 833 | eastern redcedar | northern red oak | 1 | 2 |
| 4 | 68 | 541 | eastern redcedar | white ash | 1 | 44 |
| 5 | 68 | 552 | eastern redcedar | honeylocust | 1 | 1 |
| 6 | 68 | 471 | eastern redcedar | eastern redbud | 1 | 4 |
| 7 | 68 | 931 | eastern redcedar | sassafras | 1 | 16 |
| 8 | 68 | 975 | eastern redcedar | slippery elm | 1 | 4 |
| 9 | 68 | 922 | eastern redcedar | black willow | 1 | 51 |
| 10 | 68 | 402 | eastern redcedar | bitternut hickory | 1 | 44 |
| 11 | 68 | 641 | eastern redcedar | Osage-orange | 2 | 61 |

**Fig. 3.1: An example of a portion of an FIA dataset.**

## WORKING WITH TABLEAU

Our team also chose to utilize Tableau to visualize our findings graphically. Tableau proved to be an effective way to display our findings and data graphically and efficiently.

The main project with Tableau was to create visual maps of tree locations based on their reported coordinates. We overlayed these coordinates on maps of different locations around the world, ranging from states in the United States to Angola in Africa.

While our exploration with Tableau was brief, it allowed us to gain a more visual insight into the data. Our next steps would be to expand on the types of visualizations we make to cover not just mapping but also more detailed and insightful graphs for previous projects such as Network Graphs and KNN.

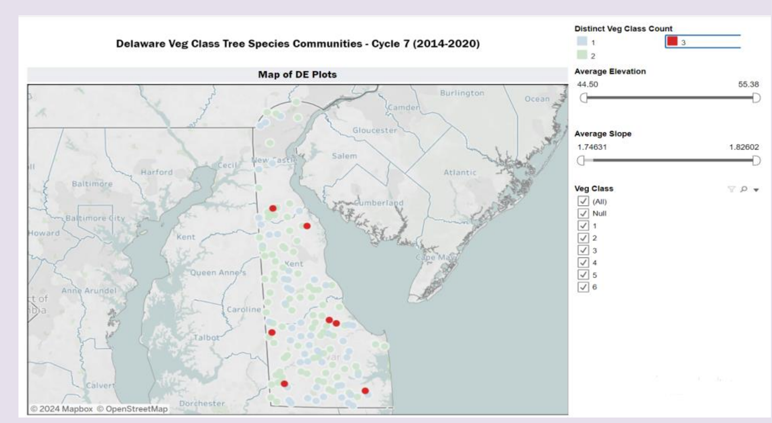


**Fig. 4: A map of Delaware's plots, focusing on the vegetation class communities. Above, the map is highlighting all the plots with a count of 3 distinct species communities.**

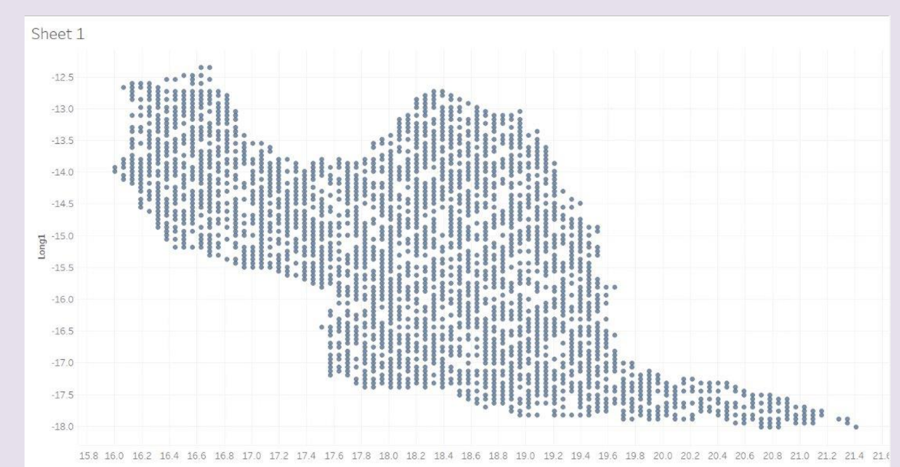


**Fig. 4.2: The known plots of Angola.**

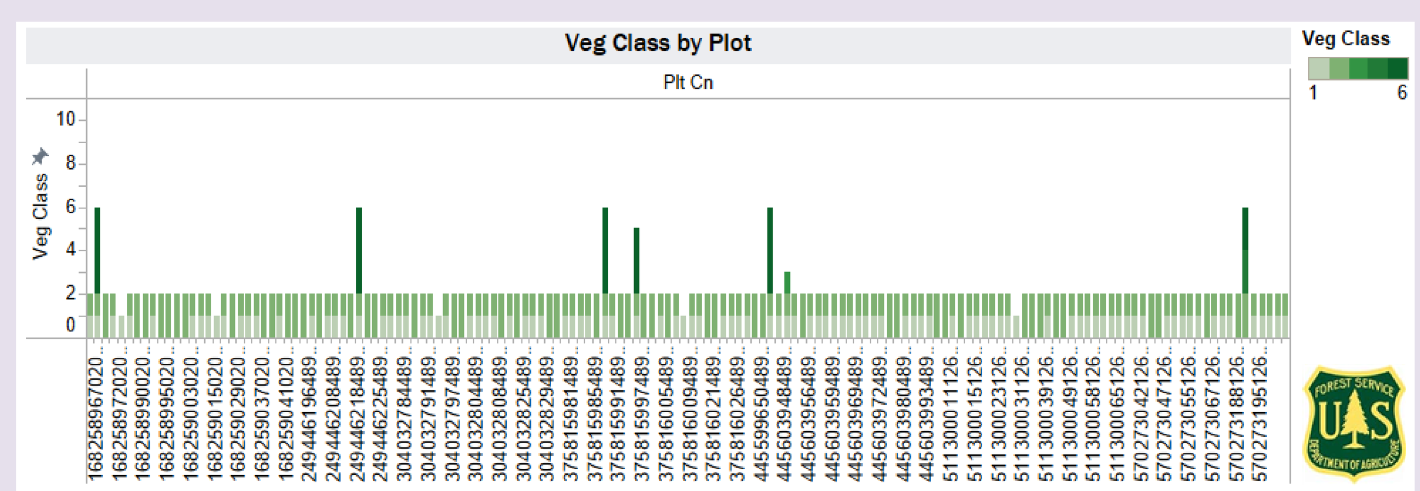


**Fig. 4.1: Shows all the vegetation class communities for each plot in Delaware.**

**Fig. 4.1.1: A QR code to an interactive Tableau Dashboard of Delaware's vegetation communities.**

## CONCLUSIONS

Throughout the course of our project, the implications of FIA data have become very clear. The database contains a wealth of information, and with enough time and knowledge, our code and methods can be used on forestry datasets of varying sizes, and locations. Moreover, our code can be used to develop models that allow for easier analysis of forest vegetation based on restrictions placed on plots.

## FUTURE WORK

- The team hopes to gather the code we have utilized in both R-Studio (which we used to create our models) and transfer it over to Tableau (data visualization software) to make model analysis easier.
- The team also hopes to replicate the models from one state to another, and possibly using it in a broader perspective (such as countries).
- Become more familiar with Tableau and explore its implications for FIA data.

## ACKNOWLEDGEMENTS


We would like to thank the Data Mine and the USDA Forest Service to allow us to work on this project. We appreciate the staff's time and effort to make this project a smooth and beneficial experience for our team.


## REFERENCES


TA's and Mentors: Andrew Lister, Rachel Riemann, Ashley Lehman, John Hogland, Mia Sartain