CS 777 Final Project

Predicting Biomass from Forest Attributes

Indigo Catton

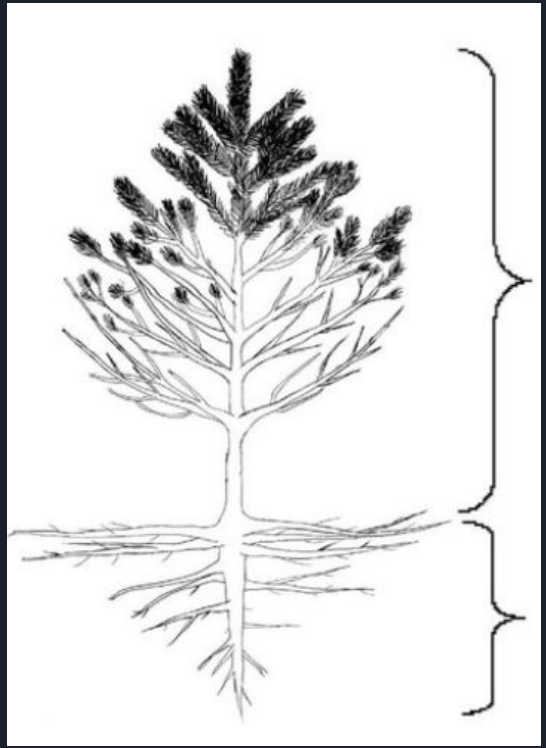
Measuring Biomass

interest in biomass for climate research and carbon sequestration

Biomass is challenging to measure and calculate accurately.

Aboveground + Belowgroud Biomass = Total Biomass

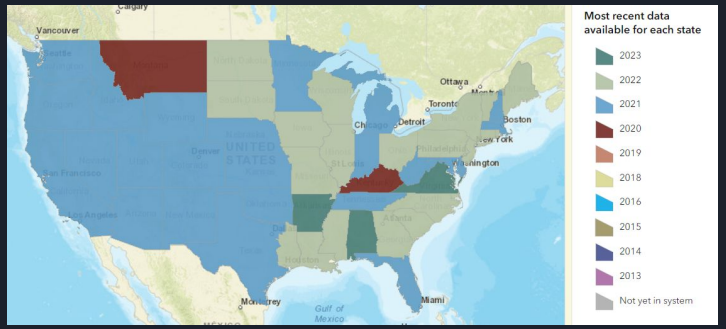
Can we use easy to measure or model variables to predict biomass?



https://www.researchgate.net/figure/Schematic-picture-of-the-traits-that-were-measured-for-the-above-and-below-ground-biomass_fig3_290473929

Forest Data

- The US Forest Service collects an annual Forest Inventory and Analysis for over 300,000 plots across the United States.
- information on trees, other vegetation, forest health, location, geography, biomass and many other features.
- Contains modeled and measured attributes
- Mix of categorical and continuous variables



FIA Data Availability

https://apps.fs.usda.gov/fia/datamart/datamart.html

Data Preparation

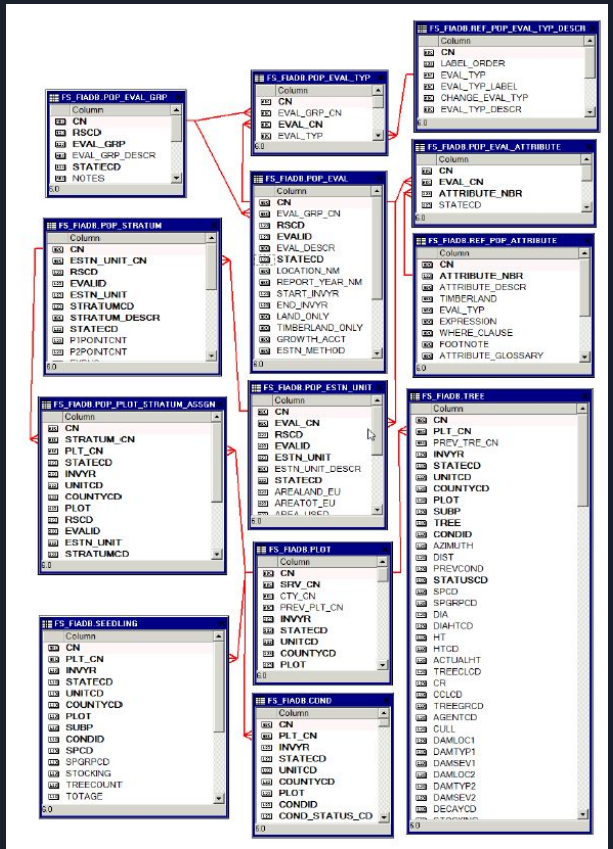
Used a subset of dataset (6,000 rows)

Data comes in a Sqlite Database

Summarized data to the plot level.

Removed sparse variables

Removed plots with null values



Database ERD

Data Preparation in Pyspark

One hot encoding on the categorical data

Used a standard scalar to center the variables

Split data 80/20 to train test

```
                "ELEV",
                "SEEDLING_TREECOUNT",
                "TREE_CT",
                "HT_SUM",
                "DIA_SUM",
                "CARBON_AG_SUM",
                "CARBON_BG_SUM",
                "CR_SUM",
                "WDLD_SUM",
                "SLOPE",
                "ASPECT",
                "STDAGE",
                "BALIVE",
                "ALSTK",
                "CARBON_DOWN_DEAD",
                "CARBON_LITTER",
                "CARBON_SOIL_ORG",
                "CARBON_UNDERSTORY_AG",
                "CARBON_UNDERSTORY_BG",
                "CONDID", #category
                "COND_STATUS_CD", #category
                "FORTYPCD", #category
                "SITECLCD" #category
```

Linear Regression

1) Linear Regression with all features
2) Linear Regression with selected features from lasso regression
3) Linear Regression with top 5 features selected with F-value

Linear Regression Results

1) Linear Regression with all features
2) Linear Regression with selected features from lasso regression
3) Linear Regression with top 5 features selected with F-value

Lasso regression did not change results, no features converged to 0 even with strong regularization.

	RMSE	R^2
All Features	3865	0.9124
Lasso	3865	0.9124
Top 5 Features	5625	0.8145

Conclusions

- Pyspark implementation allows for easy scaling.
- Initial results suggest there is potential for estimating biomass with just five easy to measure attributes.
- More work is necessary to explore the full set of available information in modeling and to better understand how to improve efficiency for large areas

Sources

https://medium.com/@agrawalsam1997/feature-selection-using-lasso-regression-10f49c973f08
https://spark.apache.org/docs/latest/ml-guide.html
https://research.fs.usda.gov/understory/forest-inventory-and-analysis-database-user-guide-nfi
Data Download: https://apps.fs.usda.gov/fia/datamart/datamart.html