FOREST GROWTH MODEL

VARSHA RAJ B, PURDUE UNIVERSITY

INTRODUCTION

Machine Learning based algorithm to estimate

- Tree Upgrowth – Change in tree width over time.
- Tree Mortality – Number of trees that have died within a defined stand over a specific period
- Stand Recruitment – Number of new trees in a forest stand.

RESOURCES

DATA : GLOBAL FOREST BIODIVERSITY – GFB3 FOR INDIANA

PROGRAMMING LANGUAGE: PYHTON

IDE: VSCODE

DATASET DEFINITION

Global Forest Biodiversity Data

- GFB3 for Indiana Tree Level Data

Number of columns: 19 | Number of rows: 122294

- PlotID, Latitude, Longitude, PA, Dmin, TreeID, Species, Status, DBH, YR, PrevDBH, PrevYR, note, Elevation_m, POM, PLT_CN, DSN_GFB3, DSN_GFB2, Orig_PlotID

Status: status code of each tallied tree

- 0 (live)
- 1 (dead)
- 2 (new)

DBH: diameter-at-breast-height (centimeters);

- DBH: 2.54 to 152.4
- PrevDBH : 2.4 to 123.4

YR: Year when the current inventory was performed;

- YR: 1996 to 2022
- PrevYR: 1984 to 2014

Distinct Plots: 4426

Distinct Species: 110

STEPS TO GET THERE

- DATA CLEANING AND PRE-PROCESSING
- SPECIES CLUSTERING AND DBH GROUPING
- CREATING SEPARATE DATASETS for M(Mortality), U(Upgrowth), R(Recruitment) MODELS
- DEVELOPING MACHINE LEARNING MODEL
- **MODEL TRAINING AND EVALUATION**

DATA CLEANING AND PREPROCESSING

1

Clean the data by removing any missing values (NA values)

2

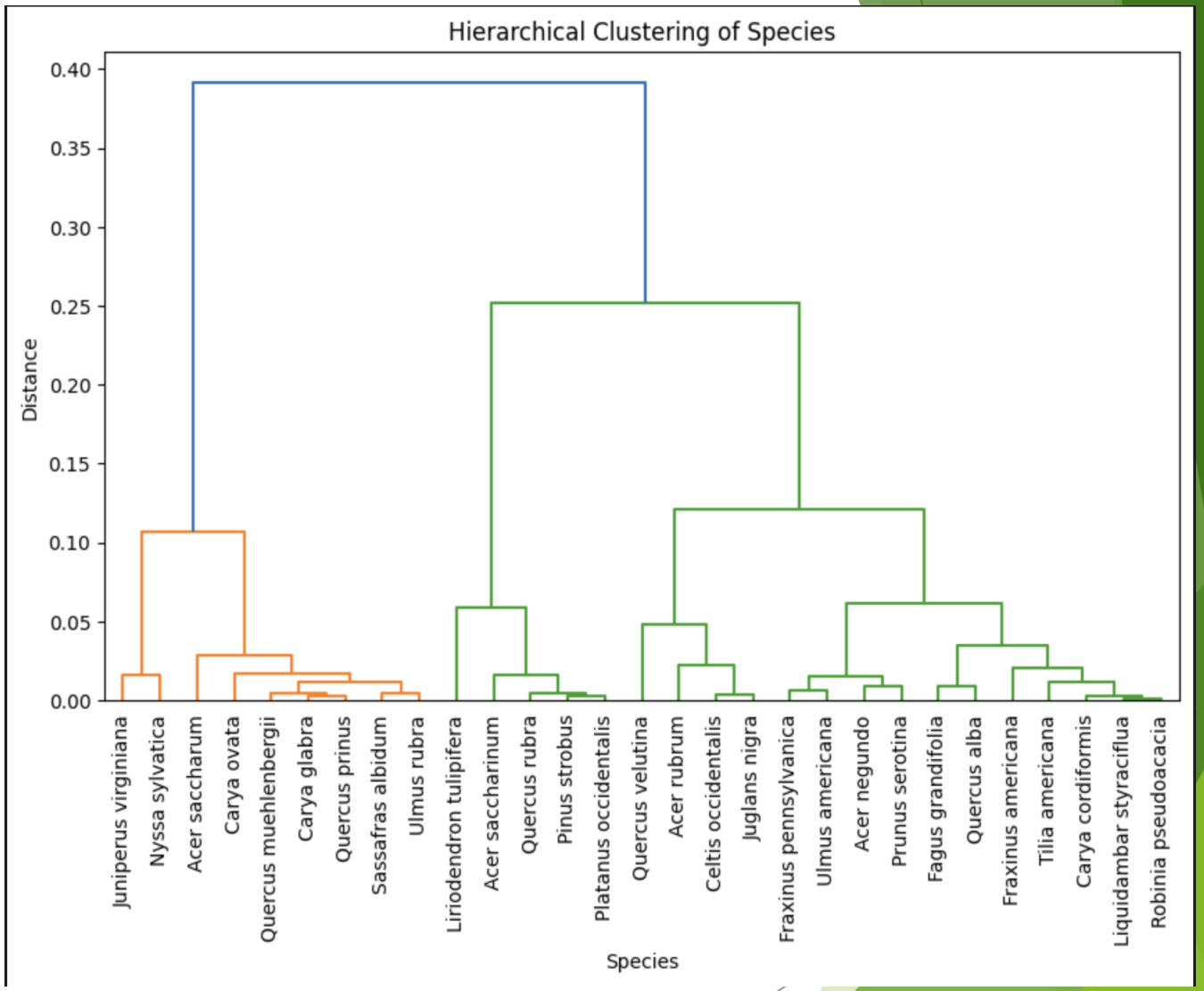
Filter the data

- Filter live trees with missing DBH & previous DBH information.
- Filter dead trees with missing previous DBH information.
- Filter new trees with missing previous DBH information.

3

Calculate necessary parameters

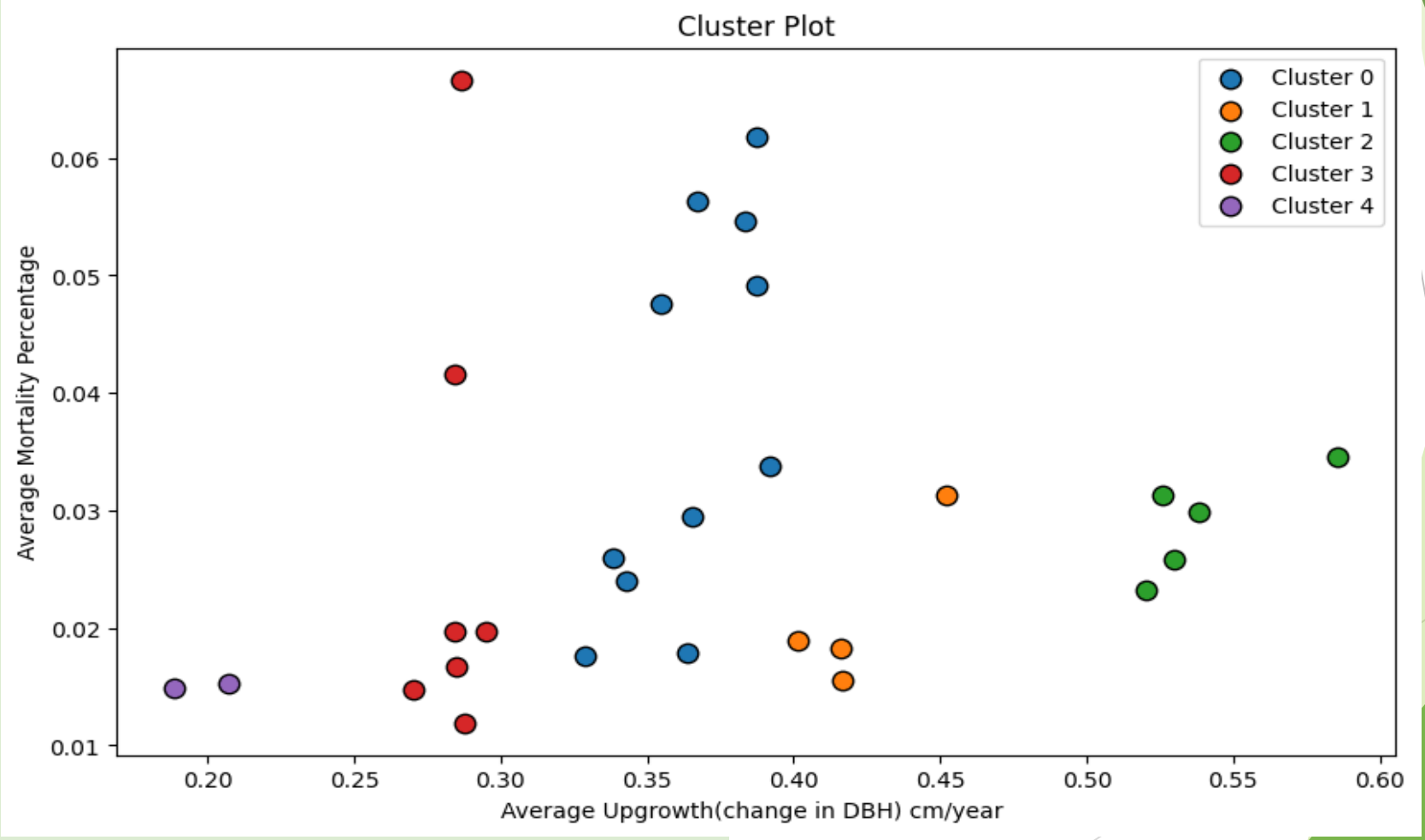
- TPH (Trees Per Hectare),
- dYR (Year Difference),
- dDBH (DBH Difference)



SPECIES CLUSTERING

- Grouping by species
  to identify significant species
  - 29 Species
- Calculating average features, such as
  - Average Upgrowth (avgDBH cm/year)
  - Average Mortality (avgMpct /year)
- Hierarchical clustering of Species
  with Euclidean distance metrics.
- Number of Clusters - 5

UPGROWTH AND MORTALITY TRENDS



SPECIES DISTRIBUTION IN CLUSTERS

```
Species in Each Cluster:
========================
Cluster 0: Acer negundo, Carya cordiformis, Fagus grandifolia, Fraxinus americana, Fraxinus pennsylvanica, Liquidambar styraciflua,
Cluster 1: Acer rubrum, Celtis occidentalis, Juglans nigra, Quercus velutina
Cluster 2: Acer saccharinum, Liriodendron tulipifera, Pinus strobus, Platanus occidentalis, Quercus rubra
Cluster 3: Acer saccharum, Carya glabra, Carya ovata, Quercus muehlenbergii, Quercus prinus, Sassafras albidum, Ulmus rubra
Cluster 4: Juniperus virginiana, Nyssa sylvatica
```

DBH GROUPING

- DBH Groups are set at intervals of 5 starting from 10 up to 60
  - 10 Groups
- Two new columns are created
  - DGP (DBH Group)
  - PrevDGP (Prev DBH Group)
- Maximum Tress fall under range (10 to 55)
  DBH: 2.54 to 152.4 PrevDBH : 2.4 to 123.4
- Feature selection and dimensionality reduction
- Capturing important patterns and relationships in the data.

Group Name	Range	DBH Count	PrevDBH Count
1	(10, 15)	14232	9812
2	(15, 20)	18985	12666
3	(20, 25)	14866	9205
4	(25, 30)	11631	6479
5	(30, 35)	8990	4592
6	(35, 40)	7760	3502
7	(40, 45)	6128	2648
8	(45, 50)	4427	1693
9	(50, 55)	3146	1106
10	(55, inf)	5924	1727
Total		96089	53430

CREATING DATASETS, ABUNDANCE MATRICES AND CROSS VALIDATION FOLDS

Creating Separate Datasets for M, U, R Models:

mortality.csv

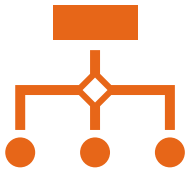
upgrowth.csv

recruitmrnt.csv

Creating Plot Abundance Matrices T1 and T2

plotsT1.csv (species occurrence across unique Plot IDs during time period 1)

PlotsT2.csv (species occurrence across unique Plot IDs during time period 2)



Creating folds for cross-validation

Random integers (1-10) are assigned to each unique PlotID

These integers serve as fold identifiers (Group1 to Group5) for cross-validation.

ML MODEL

Model Initialization: Three RandomForestRegressor models are initialized

- m for mortality prediction
- u for upgrowth prediction
- r for recruitment prediction
- Hyperparameters like the number of estimators, maximum features, and criterion ('squared_error') are specified for each model

Cross-Validation (CV):

- Cross-validation used through the nested loop.
- The data splitting is done within a nested loop that iterates over folds (from 1 to 4) and plot groups (from 1 to 9) within each fold.
- This effectively creates multiple train-test splits.

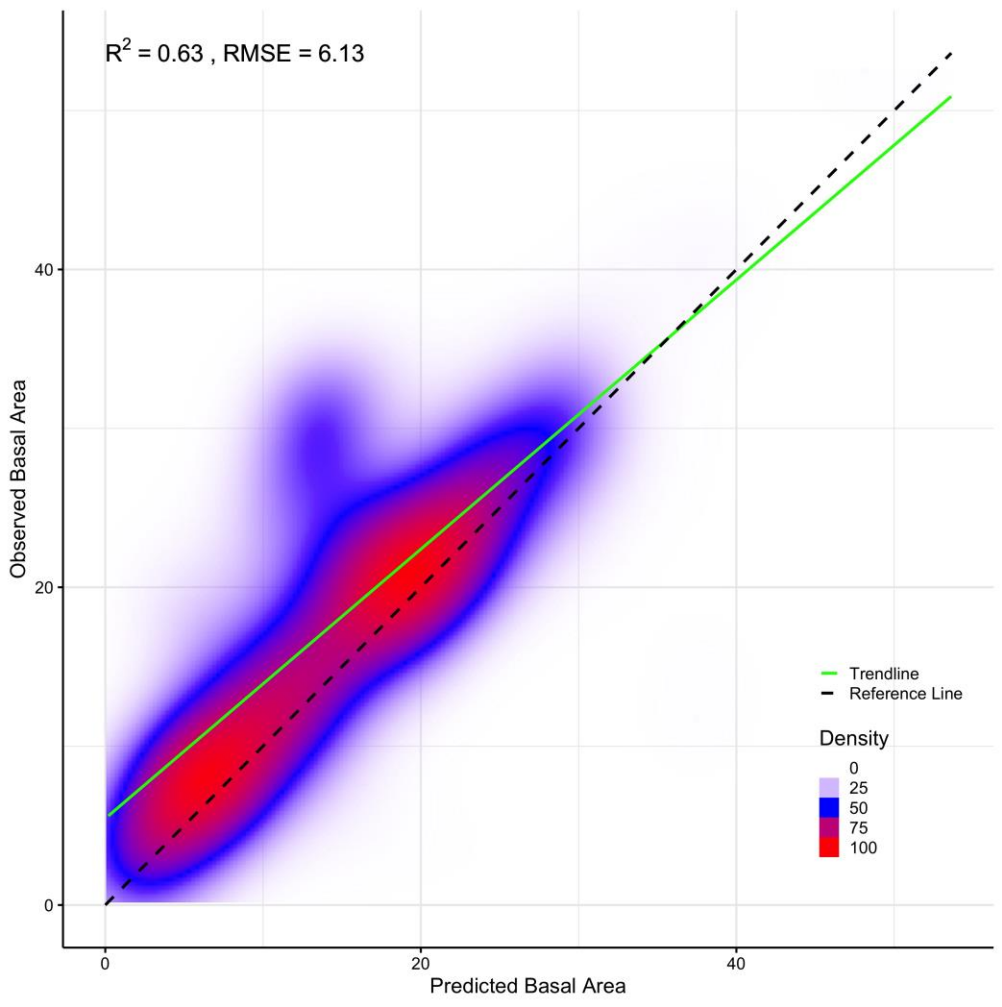
Model Training:

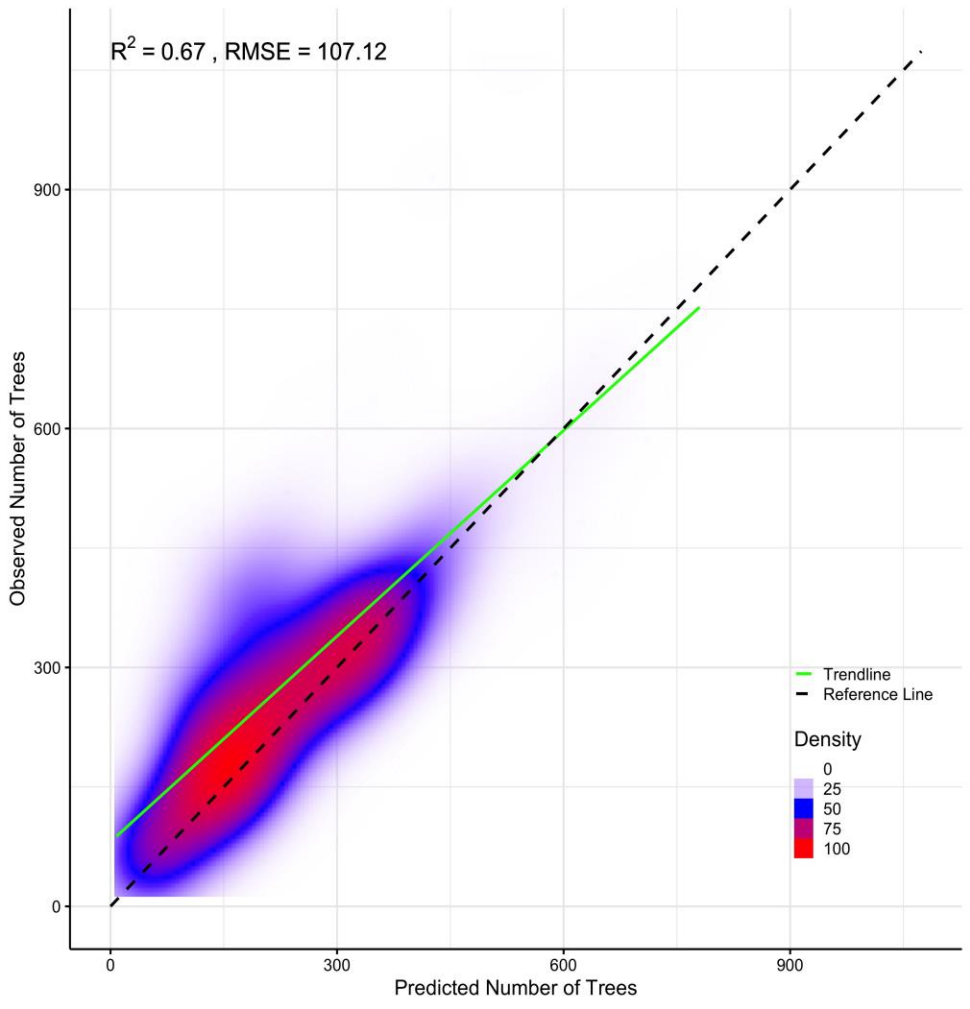
- trainm, trainu, and trainr are created which contain the relevant data needed for training.
- The initialized models (m, u, r) are trained using the (trainm, trainu, trainr) by providing the predictor variables and the target variables.

Model Prediction and Evaluation

- Calculate predictions ('mort', 'up', 'rec') using the trained models (m, u, r).
- Basal Area B and Number of trees B calculation using predicted values for T1 dataset.
- Compare with true values for B and N from T2 dataset and calculate Root mean squared error (RMSE) and R-squared (R^2)

MODEL EVALUATION





FUTURE STEPS

- Improve the accuracy and performance of the model by including covariates (Climate and Topographic Variables)
- Try out other ML models such as decision trees, or support vector machines (SVM), XGBoost to compare the results.
- Resue the code to build the forest growth model for other regions globally.

THANK YOU

VARSHA RAJ B

vbasavar@purdue.edu