Invasive Species Prediction

Bharath Sivaram, sivar019@umn.edu, Robotics MS
Pranav Julakanti, julak004@umn.edu, Robotics MS

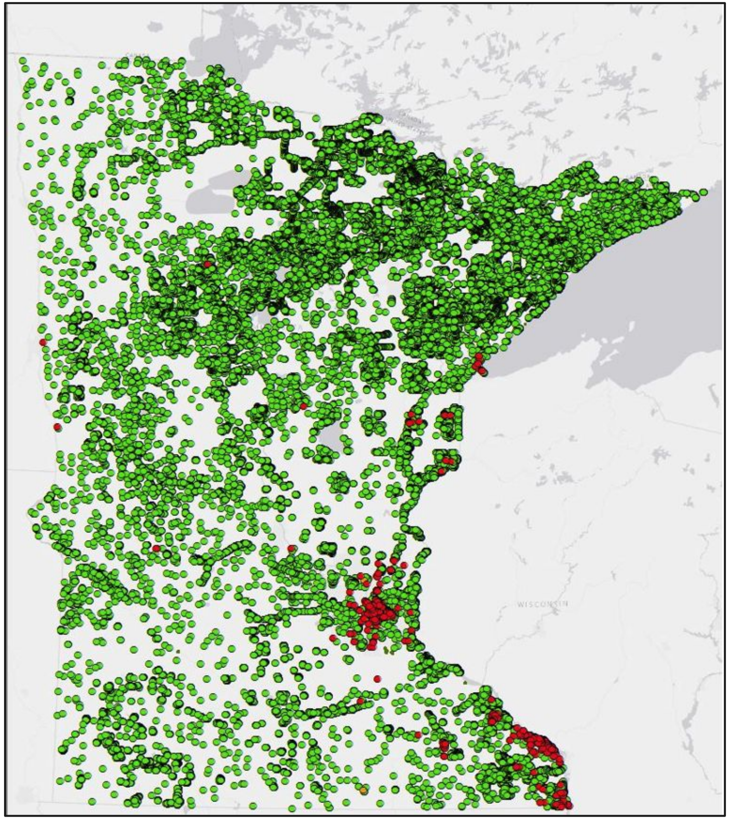
Goal

- Use species observation data to predict the certain species that will dominate in a buffer during a certain season

Canada Thistle [1]

Common Tansy [1]

Grecian Foxglove [1]



MN Terrestrial Invasive Species Observations [2]

[1]https://extension.umn.edu/invasive-species/identify-invasive-species
[2]https://gisdata.mn.gov/dataset/env-invasive-terrestrial-obs

Dataset

- We will use a gpkg provided by the Minnesota Natural Resources Department
- The dataset includes observations from 2016 to 2021
- Various features such as habitat, body type and treatment status are already included

unittype: Unit Type: Type of state land.

unitname: Unit Name: Name of the state land.

locality: Locality Description: Description about the location of observation.

site: Site Name: Specific name of area by organization.

habitat: Habitat: Area type where the subject was located.

waterbodyname: Water Body Name: Name of waterbody where subject was observed.

waterbodytype: Water Body Type: Type of water body for aquatic observations.

lakeidnumber: Lake ID Number: Lake ID number (formerly called DOW number) of the waterbody

comments: Comments: Anything that is relevant to the subject, environment, mapping.

abundance: Abundance: Distribution pattern and amount of plants, e.g. Single plant, Scattered pla

infestedareainacres: Infested Area in Acres: Actual amount of infested area within the gross area.

grossareainacres: Gross Area in Acres: Entire are a that a large or discontinuous infestation covers.

percentcover: Percent Invasive Cover: Percent cover of invasive species.

density: Density: Number of plants, or abundant, common, rare, etc.

quantity: Quantity: Number of subjects observed.

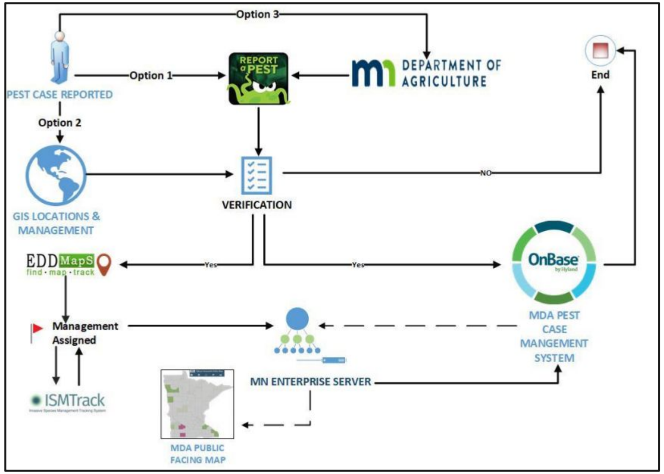
Importance

- The current system is highly dependent on reported cases
- “ We can minimize the impact of newly arrived invasive species through **early detection and rapid response” - MN DNR**
- **Our solution**

Create Model

Make Buffers within Neighborhoods/districts

Notify residents/public to keep-eye out for predicted species

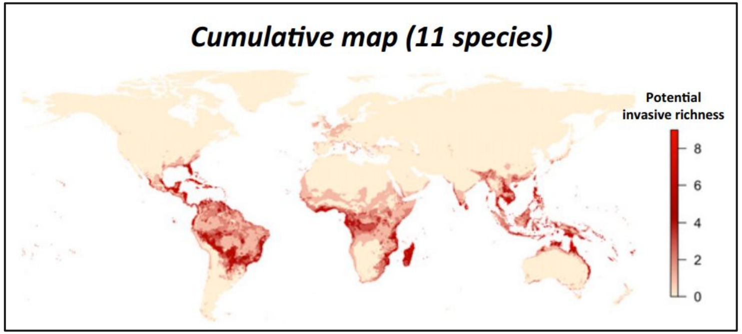


Flowchart of Pest Case Management System [3]

[3] https://www.mda.state.mn.us/sites/default/files/docs/2022-04/noxious-weed-report-2021.pdf

Domain Expert Methods

- Most work is on **distribution** of invasive species rather than predicting the species itself
- CLIMEX is commercially available modelling method [5]
  - Takes info about a species in its native range
  - Combines with climate and habitat data of area to calculate probability of invasion
- Example [4]
  - Ants favor tropical climates
  - Some ants have special trait, “supercolonial”
    Which makes them more likely to invade area
- Can also be dependent on what species already exist in area

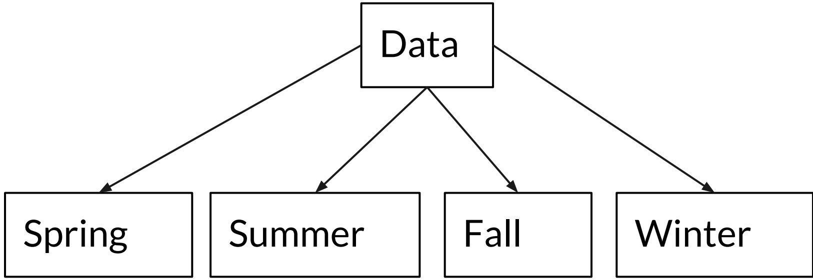


[4] Fournier A, Penone C, Pennino MG, Courchamp F. Predicting future invaders and future invasions. Proc Natl Acad Sci U S A. 2019 Apr 16;116(16):7905-7910. doi: 10.1073/pnas.1803456116
[5] Elith, J. (2017). Predicting Distributions of Invasive Species. In A. Robinson, T. Walshe, M. Burgman, & M. Nunn (Eds.), *Invasive Species: Risk Assessment and Management* (pp. 93-129). Cambridge: Cambridge University Press. doi:10.1017/9781139019606.006

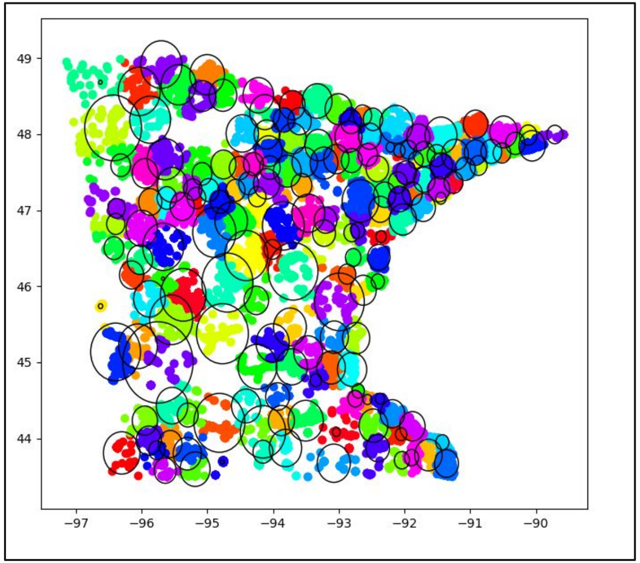
Approaches and Assumptions

- We will use a random forest and NN
- Currently, the most common variables used are climate and precipitation in addition to location/habitat
- We make an assumption that by splitting up by season, we account for climate/precip
- We also assume that OSM features vectors capture details about location/habitat
    - For example, the “natural” feature has details such as grassland, woods,water, etc.

Data Pre-Processing



- Split data up by season, based on equinox/solstice [x]
- First attempt to create buffers was using clustering, but resulted in too large buffers (points are spread)
- Use PostGres to create buffers around each data point
  - Under the assumption that similar environments will have similar species
- Find overlaps of buffer and OSM features to create geo-feature vectors
- We concatenate OSM features with the habitat features from the Minnesota Dataset

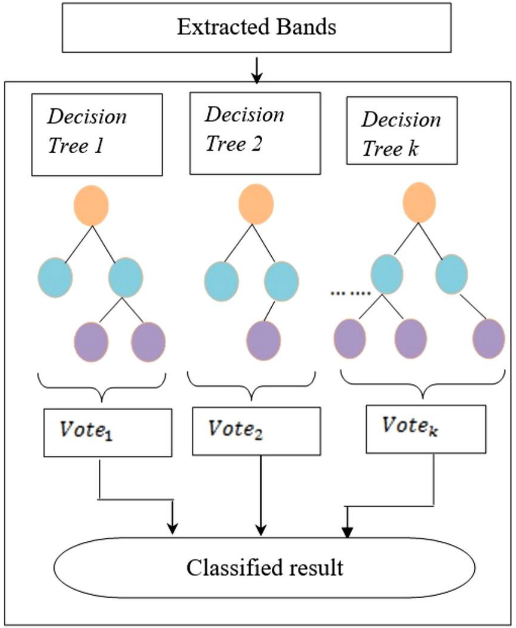


Original Buffer Attempt

[x] https://www.timeanddate.com/calendar/seasons.html?n=159

Random Forest Approach

- We assign a numeric label to each species
- A random forest model is trained on a training subset of the data
- After testing, we found performance peaks at 100 trees.
- A validation subset is used to assess performance



Random Forest Results

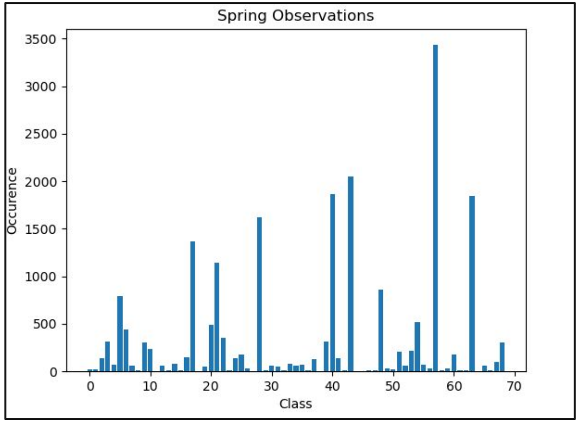
- 66% accuracy
- Many natural features are important
- The data was dominated by a few labels, many feature vectors were sparse

Feature	Importance
natural_water_1000	0.04728141010919247
highway_service_1000	0.046471756984601806
highway_residential_1000	0.04064659862631857
natural_water_500	0.03905779941179456
waterway_stream_1000	0.037728277101964476
highway_service_500	0.03578585560644963
highway_residential_500	0.03532595356125794
landuse_retail_1500	0.030976671060348987
waterway_stream_500	0.02905472393657684
highway_primary_1500	0.027542490185913097
highway_service_1500	0.023879037820748683
highway_secondary_link_1500	0.020881306108315654
highway_secondary_500	0.02052002145470157
natural_wood_1500	0.018543424096022516
waterway_river_1000	0.01815385584398259
landuse_recreation_ground_1500	0.017879371282333936
highway_footway_1000	0.015266951407040815
highway_unclassified_1000	0.01483645817780274
waterway_river_500	0.014618784444123685
highway_track_1500	0.013705767748889536
waterway_ditch_1000	0.013460702667722731
highway_track_500	0.013302411101726265
landuse_forest_1000	0.012959849868325513
amenity_parking_1000	0.011487174032524021
highway_residential_1500	0.010933815796672835
landuse_quarry_1500	0.010822149287245127

NN Sampling & Overview

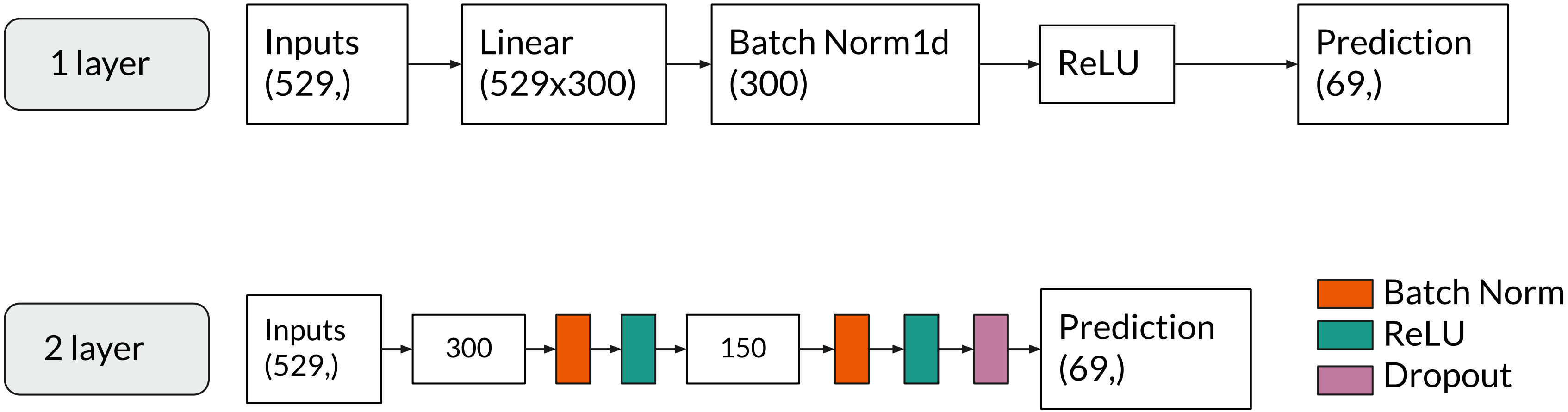
- Classes were heavily imbalanced, which required attention when splitting into train,test,val
  - Used stratification to get same ratios of classes in train & val
  - Also used WeightedRandomSampler where each sample weight is inverse of the frequency of the class associated with sample
- Tried 2 different network architectures
  - 1 layer, # neurons = mean(input,output)
  - 2 layer, incorporating dropout [x]
- 2 different optimizers while testing multiple learning rates
  - Adam
  - SGD

Class Distribution



[x] https://towardsdatascience.com/pytorch-tabular-multiclass-classification-9f8211a123ab

NN Architectures



[x] https://towardsdatascience.com/batch-norm-explained-visually-how-it-works-and-why-neural-networks-need-it-b18919692739

[x] https://stats.stackexchange.com/questions/181/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw

NN Results

Training Condition	Spring Loss	Spring Acc (4212 testcases)	Summer Loss	Summer Acc (16313 testcases)
1 layer, SGD,	0.59	0.396	0.82	0.424
2 layer, SGD	2.18	0.25	0.75	0.477
1 layer, Adam,lr0.01	0.32	0.394	0.77	0.521
2 layer, Adam,lr0.01	0.22	0.425	0.45	0.583
2 layer, Adam, no dropout	0.58	0.396	0.67	0.446

Note: We are deciding between 69 classes for spring and 85 classes for summer

NN Results

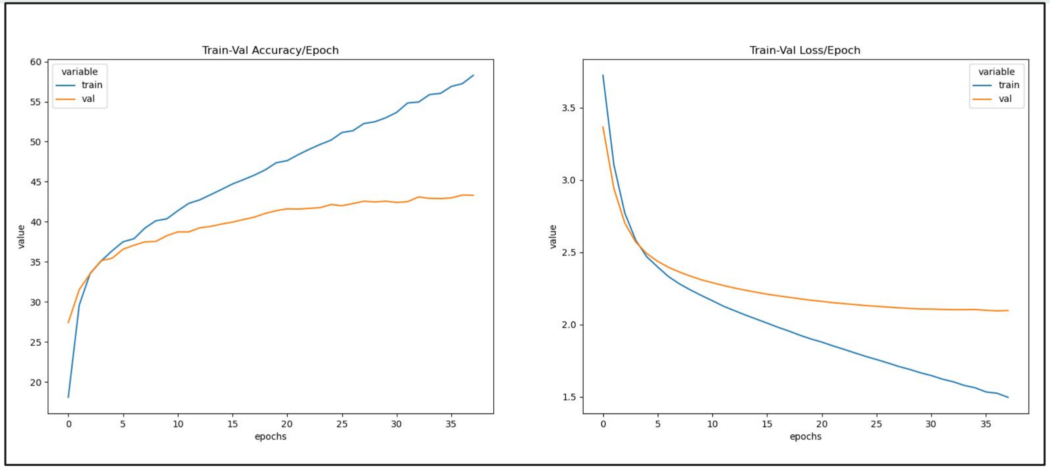
- Validation losses never got below 2 for any of our cases

-This is likely due to overfitting to training

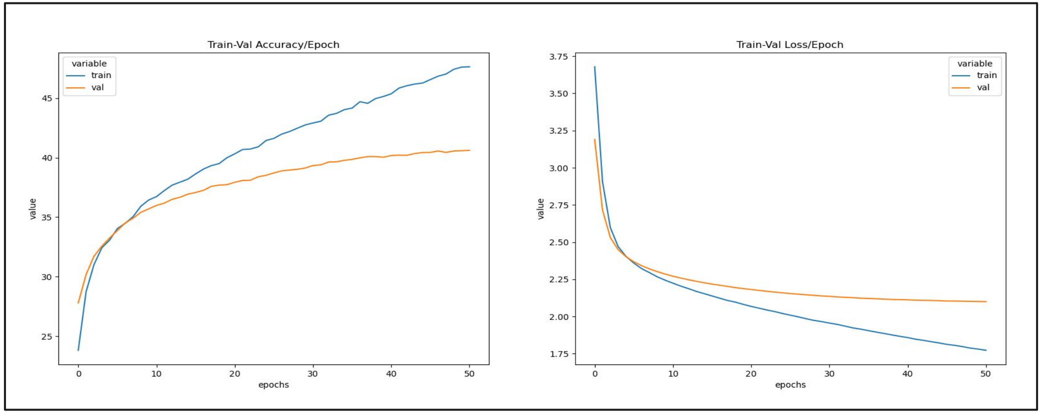
-Need to take closer look at hyper params:

- Batch Size
- Learning Rate
- The exact # of neurons within each layer

-The data just may not be complex enough to warrant a NN, XGboost Might be a better approach



Spring Data Training performance



Summer Data Training performance

Challenges and Future Work

- The observation points are sparse, the dataset isn't expansive for the state of minnesota
- Most of the observations have sparse OSM features. OSM categories work better for urban environments
- The data from the Minnesota Natural Resources Department has many missing fields. The data had little impact on our algorithms
- Class Overrepresentation.
  - This is a non-avoidable problem since some species are just more dominant
- Two possible steps forward
  - Incorporate more expert opinion/data
    - Average precipitation/temp by month
    - Find better ecological data factoring in expert assumptions
  - Use satellite images for prediction rather than raw geographic info

Questions or Comments?