Minnesota Invasive Species Prediction by Season

Bharath Sivaram
Pranav Julakanti

Introduction

Invasive species are one of the largest problems within ecological systems. A 2021 study found that they have cost North America close to $26 billion per year since 2010 **[1]**. They also wreak havoc on native populations due to predation and disease. In this paper, we focus on invasive species in Minnesota, which is under severe risk due to the number of lakes in the state. In fact, the zebra mussel has become the poster child of Minnesota invasive species due to its prevalence in lakes.

Although areas of species are known, treatment is a challenge on its own, since chemical treatment is not only high-cost in dollar amount, but can also impact surrounding species **[2]**. For that reason, total eradication is usually never an option, rather, a long-term plan is established for reduction. This is both resource and time intensive. The current model also depends on individual observation reports for tracking of species over time, which costs the State money and can be incomplete since it's difficult to get to all areas. The MN DNR states in its 2021 Annual Report, "We can minimize the impact of newly arrived species through early detection and rapid response."[**3**]. This paper describes a model which allows for prediction of a dominant species within certain buffers zones in the states. The aim is to use these buffer predictions to inform the public what species they should be alert for and report based on their residence. This will solve two issues: 1) People don't know what to look for, and 2) By the time people notice an infestation, it is too late. We hope this method results in more info for the state while avoiding the cost of sending employees out to do observations.

Related Work

The majority of work done in this area is focused on invasive species distributions rather than prediction. However, there has been fundamental research within the past 20 years which mainly focus on mathematical and phenological models. One example is CLIMEX, which is a commercially available modeling method created in the 1980s and used to this day by many government agencies. CLIMEX takes into account the climate and habitat data of grid spaces in combination with phenology details of species within their natural habitat **[4]**. They then predict the probability of a species invading that grid space. In addition, the model can also predict the growth rate of that species based on factors such as temperature, moisture, light, etc. One example paper uses this methodology to predict possible areas of invasion for ants **[5]**. Specifically, they use phenology and behavior in native habitats to classify different species as "super colonial", meaning they are more prone to domination over other species. Additionally, the climate and precipitation data from different regions is used since ants favor tropical climates. The result is a cumulative map where a color-scale is used to indicate "potential invasiveness". Although this method can be accurate, the main problem is that it is not possible to go through every species and do this for Minnesota. The method also requires deep knowledge about the biological traits of each species, which is not always available. CLIMEX also requires payment and experts.

The only other current practice in the state is sending out DNR workers to take observations in select locations. This means many grid spaces will be left unchecked and ripe for infestation. It also takes up working time which could be used for other tasks within the DNR. This is why we use machine learning combined with observation data till current to solve the problem.

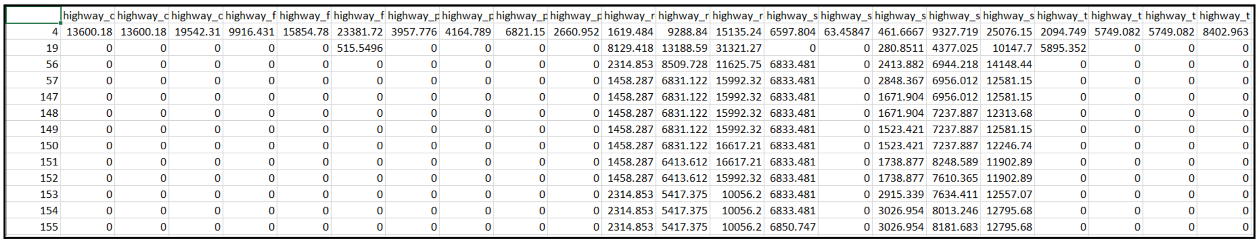
Approach

This paper describes two attempts to solve the issue. The MN Terrestrial Invasive Species Observations dataset was used for both attempts **[6]**. This dataset has species observations made within the last 5 years within the state of Minnesota. The locations, species name, observation dates, and habitat info are included in this data. Based on related work, we know the most important factors for species prediction is climate/precipitation and habitat. To account for climate/precipitation, we pre-process data by splitting up by season based on the equinox/solstice dates in Minnesota. We also use OpenStreetMap (OSM) features with the assumption that they can account for habitat information. OSM has natural features such as shrubbery, wood, wetland, etc. In fact, due to incomplete habitat data in the MN dataset, OSM turns out to be the main habitat information provider. After data is split into seasons, buffers of 500,1000, and 1500 meters are created around each point. Then the OSM features within each buffer are accumulated to create a feature vector for each observation point. This was done under the assumption that clumped points will have similar features, and therefore the prediction will result in the densest species. This approach of buffers and feature vectors follows the methodology in a paper for air quality prediction using public datasets **[7]**. Once these feature vectors are made, the species names are turned into a classification vector where each name is associated with an integer class. Random forest (RF) was the first approach, and this was followed by an artificial neural network (ANN). Due to observation data lacking in the fall and winter, we focus on results in the spring and summer data in this report. The goal will be to predict the species within a buffer based on the features within that buffer. By using OSM, we have a large amount of location data that can be used to characterize neighborhoods and other residential areas. Should the prediction accuracy be 80%+, the model can be a tool for the DNR to alert the public regarding certain species, and also to send out workers for certain areas of interest based on the predicted species.

Experiments

Random Forest

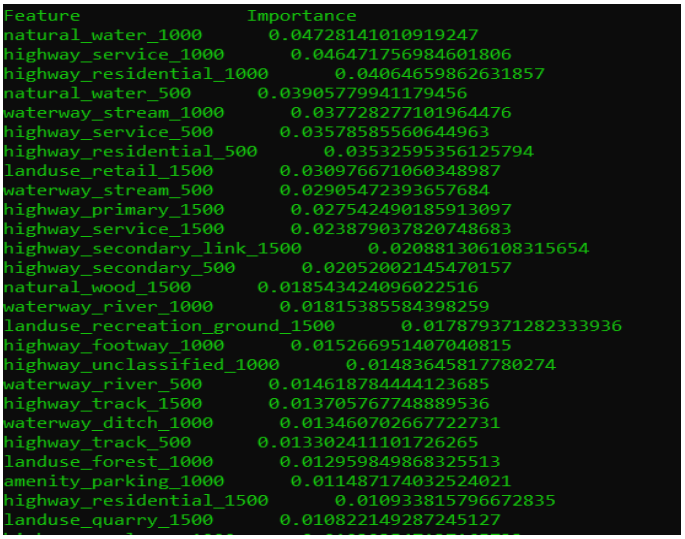
We used Random forest to classify species. The goal of random forest is to use geographic features as predictors for species observations. We use the OSM dataset with some ecological features to create vectors which describe the environment surrounding a species. An example of this vector is shown in Figure 1 below.



	highway_c	highway_c	highway_c	highway_f	highway_f	highway_f	highway_p	highway_p	highway_p	highway_p	highway_r	highway_r	highway_r	highway_s	highway_s	highway_s	highway_s	highway_s	highway_t	highway_t	highway_t	highway_t
4	13600.18	13600.18	19542.31	9916.431	15854.78	23381.72	3957.776	4164.789	6821.15	2660.952	1619.484	9288.84	15135.24	6597.804	63.45847	461.6667	9327.719	25076.15	2094.749	5749.082	5749.082	8402.963
19	0	0	0	0	0	515.5496	0	0	0	0	8129.418	13188.59	31321.27	0	0	280.8511	4377.025	10147.7	5895.352	0	0	0
56	0	0	0	0	0	0	0	0	0	0	2314.853	8509.728	11625.75	6833.481	0	2413.882	6944.218	14148.44	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	1458.287	6831.122	15992.32	6833.481	0	2848.367	6956.012	12581.15	0	0	0	0
147	0	0	0	0	0	0	0	0	0	0	1458.287	6831.122	15992.32	6833.481	0	1671.904	6956.012	12581.15	0	0	0	0
148	0	0	0	0	0	0	0	0	0	0	1458.287	6831.122	15992.32	6833.481	0	1671.904	7237.887	12313.68	0	0	0	0
149	0	0	0	0	0	0	0	0	0	0	1458.287	6831.122	15992.32	6833.481	0	1523.421	7237.887	12581.15	0	0	0	0
150	0	0	0	0	0	0	0	0	0	0	1458.287	6831.122	16617.21	6833.481	0	1523.421	7237.887	12246.74	0	0	0	0
151	0	0	0	0	0	0	0	0	0	0	1458.287	6413.612	16617.21	6833.481	0	1738.877	8248.589	11902.89	0	0	0	0
152	0	0	0	0	0	0	0	0	0	0	1458.287	6413.612	15992.32	6833.481	0	1738.877	7610.365	11902.89	0	0	0	0
153	0	0	0	0	0	0	0	0	0	0	2314.853	5417.375	10056.2	6833.481	0	2915.339	7634.411	12557.07	0	0	0	0
154	0	0	0	0	0	0	0	0	0	0	2314.853	5417.375	10056.2	6833.481	0	3026.954	8013.246	12795.68	0	0	0	0
155	0	0	0	0	0	0	0	0	0	0	2314.853	5417.375	10056.2	6850.747	0	3026.954	8181.683	12795.68	0	0	0	0

Figure 1. Feature Vector

After creating features vectors, the data is normalized and fed into a random forest network. We use 100 trees. Random forest prevents overfitting from the fact that each tree uses a subset of the features. We split the dataset into training and validation subsets. Since the dataset does not have a history of observations like the pm25 dataset, we were unable to use cluster analysis tools such as k-means to determine important features. Regardless, we were able to achieve an **accuracy of 66% and 71%** with the spring and summer datasets respectively. **Figure 2** below shows the most important features used by the random forest algorithm.



```
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
```

Figure 2. Feature importance scores

As we can see, natural classes such as “wateway_streams”, “natural_water”, “”landuse_forest” and others appear as expected. Surprisingly, the algorithm identifies highways as an important predictor.

ANN

The two ANN architectures are shown in **Figure 3**. Batch norms are used to speed up training. For the 1 layer network, the hidden layer of 300 is chosen as the mean of the input and output vectors.

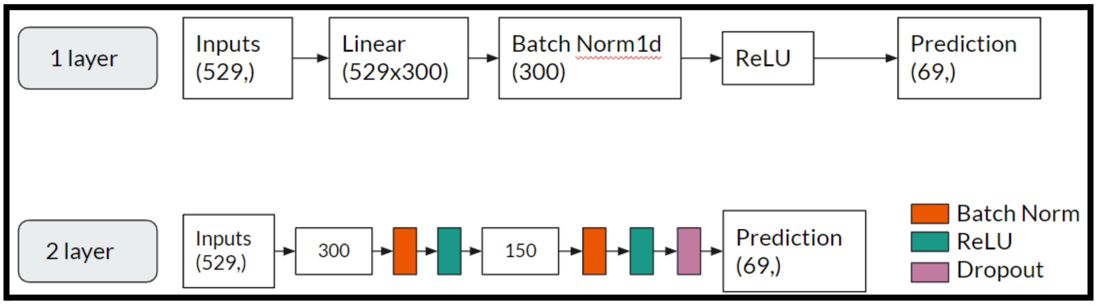


Figure 3. NN Architectures

The main issue in data was large skew between observations, where a few species dominated, as shown in the histogram in **Figure 4**. To mitigate this, stratification was used when splitting data into train,val,test sets. This ensures the same ratio of classes exists in all sets, in other words, we wouldn’t have a case that a completely new species appears in test/val that wasn’t in training. A WeightedRandomSampler was also used, allowing rare classes to be sampled more in training than the overrepresented classes. It’s implemented by a weight vector with lower weights for majority class samples. Both Adam and SGD optimizers were tested, with a learning rate of 0.01. Due to lack of observations, spring and summer data were the main focus. A split of 80-20 was used for training-testing, and CrossEntropyLoss was used for training. The results of the ANNs are shown in Table 1.

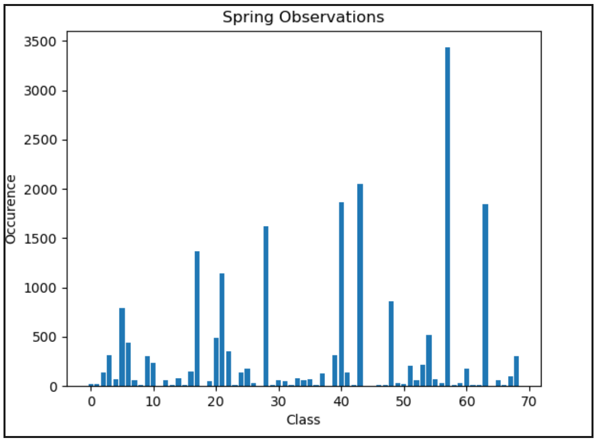


Figure 4. Spring Observations Distribution

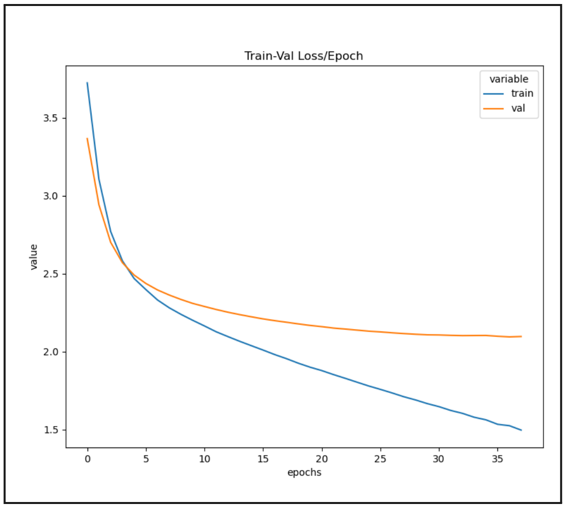


Figure 5. Spring Data Training Performance

Table 1

Training Condition	Spring Loss	Spring Acc (69 classes)	Summer Loss	Summer Acc (85 classes)
1 layer, SGD,lr=0.01	0.59	0.396	0.82	0.424
2 layer, SGD,lr=0.01	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

Discussion

The NN approach results were underwhelming when compared to the random forest approach, due to lower accuracies for all training methods. The likely reason for this is overfit to training data. As shown in **Figure 5**, by epoch 5, the model starts to overfit and minimal improvement is made to validation loss. Additionally, the losses never drop lower than 2. It is important to note however that the model is deciding between 69 and 85 classes. While the Random Forest approach performed better, it faced many of the same issues with the Neural Network approach. Firstly, the features we used to train the network were sparse. OSM only has a handful of ecological classes. The network struggled to classify species using such sparse data. Additionally, the dataset had a large imbalance of classes. While some species appear in the dataset for hundreds of observations, others were recorded less than 10 times. The dominance of these labels made it difficult for the algorithm to perform well with the underrepresented labels. Finally the dataset provided by the state of Minnesota was poorly maintained. Many observations had missing fields for “habitat” and “treatment status”.

Future Work

In future, it would be ideal to have a more complete dataset while also taking advantage of temp/precipitation data when training/predicting. This would mitigate two of the largest issues , which are the spread of observations compared to the size of Minnesota being quite sparse, and the overrepresentation of classes. Additionally, it would be worthwhile to predict using satellite imagery rather than geographic features since images can have many details about habitat encoded in them.

Incorporating expert opinions such as species biology and preferences would also be helpful since they are known to have an impact on where they tend to invade.
As for training changes, it would be better to split data based on location rather than randomly. This way, the model can be tested on a location it hasn't seen before since clumped points can skew the testing losses to look better than they are. And finally, XGBoost should be implemented in future since the data has shown stronger performance on the tree-based model.

Sources

[1] *Economic and social impacts*. Economic and Social Impacts | National Invasive Species Information Center. (n.d.). Retrieved May 5, 2022, from https://www.invasivespeciesinfo.gov/subject/economic-and-social-impacts

[2] *Invasive Aquatic Plant Management (IAPM)*. Minnesota Department of Natural Resources. (n.d.). Retrieved May 5, 2022, from https://www.dnr.state.mn.us/invasives/iapm.html

[3] *2022 noxious weed list - mda.state.mn.us*. (n.d.). Retrieved May 5, 2022, from www.mda.state.mn.us/sites/default/files/docs/2022-02/2022NoxiousWeedListFactsheet.pdf

[4] 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
[5] 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

[6] *Terrestrial Invasive Species Observations*. Minnesota Geospatial Commons. (n.d.). Retrieved May 5, 2022, from https://gisdata.mn.gov/dataset/env-invasive-terrestrial-obs

[7] Lin, Y., Chiang, Y.-Y., Pan, F., Stripelis, D., Ambite, J. L., Eckel, S. P., & Habre, R. (2017). Mining public datasets for modeling intra-city PM2.5 concentrations at a fine spatial resolution. *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. https://doi.org/10.1145/3139958.3140013