Programming For Scientists Project Report - Invasive Species Modeling

Jonathan Zhu

2023-10-30

Background

Amidst growing concerns about the environment involving climate change, eutrophication, deforestation, and other human impacts comes not only the loss of biodiversity through human actions but also the spread of invasive species. Now present in all parts of the world, from marine environments to mountainsides, invasive species are any species that have migrated through some means to a new environment where they would not normally be found without outside intervention.

Invasive species tend to have negative effects on the ecosystems in which they are introduced; since they have no natural predators in the new environment, they have a clear competitive advantage and can grow however they like, often consuming large amounts of nutrients and harming other organisms. This in turn leads to loss of biodiversity and thus ecosystem degradation. While us as modern-day humans are more disconnected from nature than ever, we still benefit greatly from it and should do what we can to mitigate the spread of invasive species. One example of this is the zebra mussel collapsing freshwater fisheries, hurting the businesses supplying those fish.

While there are many ways to curb the spread of invasive species, from cleaning boats to cleaning shoes, total eradication of these species is near-impossible thanks to their prevalence, rapid grow rate, and general lack of resources, especially funding. While a few exceptions exist, invasive species control is generally more focused on preventing the spread of these species rather than eradication. As such, it is important to understand how invasive species spread, and one tool for that is the computational power of today.

However, modeling invasive species spread poses a myriad of difficulties. For one, there are many factors involved; these range from abiotic factors such as climate and soil condition to human factors such as boat traffic and population density. Mainali et al. (2015) is one paper that attempts to quantify some of these factors into modeling the spread of invasive species, and I attempt to recreate their model in this project, with a few twists, which are described later.

In essence, the work of Mainali et al. uses climactic variables provided by `worldclim` (Table 1), part of R's `raster` package, alongside latitude, longitude, elevation, depth, and a few anthropogenic variables to predict the spread of the *Parthenium* weed. They subsequently recruit the help of machine learning models, namely the pre-existing Generalized Linear Model (GLM), random forest, and boosted forest. This leads to the first of my additions, which is the use of alternative methods for model construction. Mainali et al. use the `biomod2` R package, and I attempt to do the same. However, I also wanted to construct additional models using the `tidymodels` suite of R packages; this is a much more accessible way for machine learning to be conducted as it does the majority of the work behind the scenes.

Table 1: Table 1: The `worldclim` climactic variables and their descriptors derived from the `worldclim` documentation.

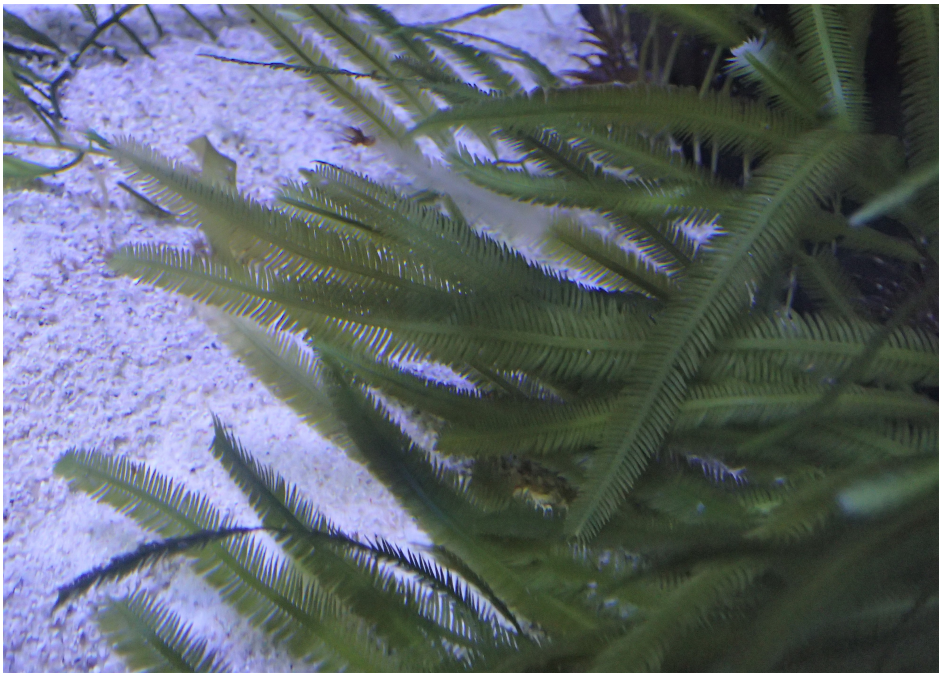
Variable Name	Description
Temp	Annual Mean Temperature
Diurnal_Range	Mean of monthly temperature range

Variable Name	Description
Isothermality	Diurnal Range / Annual Range
Seasonality	Standard Deviation of Annual Temp
Max_Temp	Max. Temperature of Warmest Month
Min_Temp	Min. Temperature of Coldest Month
Wet_Temp	Mean. Temp. of Wettest Quarter
Dry_Temp	Mean. Temp. of Driest Quarter
Annual_Range	Max Temp. - min. temp.
Warm_Temp	Mean Temp. of Warmest Quarter
Cold_Temp	Mean Temp. of Coldest Quarter
Prec	Annual Precipitation
Wet_Prec	Precipitation of Wettest Month
Dry_Prec	Precipitation of Driest Month
Wet_Prec_Quart	Precipitation of Wettest Quarter
Dry_Prec_Quart	Precipitation of Driest Quarter
Prec_Seasonality	Coefficient of Variation for Prec
Warm_Prec	Precipitation of Warmest Quarter
Cold_Prec	Precipitation of Coldest Quarter

The second of my additions to the work is the addition of more species. The subject of choice for Mainali et al. was the *Parthenium* weed, commonly known as feverfew (shown directly below). Native to North America, it has a myriad of human uses from a painkiller to a natural rubber source; however, it is an invasive species throughout India, Australia, and Africa. The *Parthenium* weed was chosen to evaluate the role of roads, population density, soil moisture, and canopy cover in the spread of invasive flora. However, I chose to extend a similar methodology of Mainali et al. to five more species, each described below.

The marine alga *Caulerpa taxifolia* (shown directly below) is an invasive plant-like multicellular protist native to various tropical oceans. It is commonly used in aquaria as decoration; however, this is the main reason for its spread, as aquaria leaks are identified as the source of its invasion in the Mediterranean Sea and California. Controlling its spread is crucial as it is known as one of the worst invasive species in the world, being able

to overtake marine ecosystems with little resistance and wiping out native flora and fauna, so I applied the procedures to the data of its occurrences.



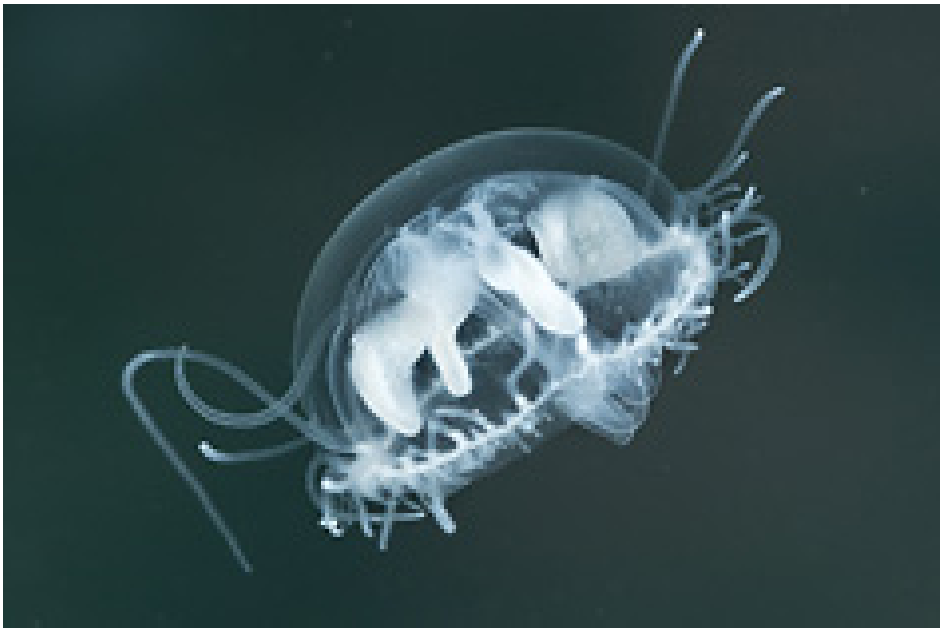
The tunicate *Ciona intestinalis* (shown directly below) is a sessile invertebrate commonly found on man-made marine structures like docks and ropes. Found in all the world's oceans, it has been an invasive species for every ecosystem in which it is observed; its native habitat is unknown. While it is a laboratory model organism and has informed us on studies concerning regeneration and aging, it is more so regarded as a pest, and so understanding its spread is crucial, so its data was also used in this project.

The zebra mussel (*Dreissena polymorpha)* is another invasive species regarded as a pest for humans and is one that is incredibly damaging to freshwater ecosystems. Native to Eastern Europe, the zebra mussel has been documented in many states in the US and is known to filter all nutrients out of a freshwater system when present in large amounts, killing everything else present in the system. This, of course, has negative

implications for freshwater fisheries; additionally, since it grows on all kinds of surfaces, its sharp shell edges form a danger to humans. Understanding its potential spread is crucial to prevent further irreparable ecosystem damage for the places where it is currently not present.

The spotted lanternfly (*Lycorma delicatula)* was a common sight for us in Pittsburgh; though native to China, we at Carnegie Mellon should know very well how it shows up in such large amounts and can easily overtake native species of insects, which again poses negative implications for fauna. While we can hope that the cold winters of Pittsburgh can kill them off, this is no guarantee; as such, it is also important to understand its spread.

Finally, I chose to run the same methods on the freshwater jellyfish *Craspedacusta sowerbii* (commonly known as the peach blossom jellyfish); while many of its implications and effects on freshwater ecosystems are unknown, I chose to do it as a passion project as my undergraduate research in my senior year was focused on this organism.



Data Retrieval and Processing

Occurrence data for the six common invasive species described in the background, those being *Parthenium sp., Caulerpa taxifolia, Ciona intestinalis, Dreissena polymorpha, Lycorma delicatula,* and *Craspedacusta sowerbii*, were obtained from the Global Biodiversity Information Facility (GBIF), one of the most comprehensive open collections of species occurrence data in the world. While Mainali et al. used these occurrence data, they used several others, including many private collections, which were not accessible and as such not used in the dataset for this project.

Predictor variables included latitude and longitude, present with the occurrence data from GBIF; elevation, depth (if applicable), and nineteen standard climate variables found in the WorldClim data loaded in R's `raster` package. These variables are commonly used for analyzing world climates and include temperature and humidity. While data on soil moisture was obtained, information on roads, population density, and canopy cover were unfortunately tied to defunct websites or currently irrelevant file formats and so were not used as predictor variables for spread. Additionally, for the other five species, I predicted these predictor variables would be far less useful in predicting spread.

Given occurrence data, I labeled all occurrences with 1.0 ("Yes") as having the species present at that point. I then created "dummy" datasets filled with any points on the planet that did not have the species present, as well as their respective WorldClim variable values. These points were artificially selected using Golang, exported to R, and then filled with their respective values. Following this, if any random points were duplicates, they were removed. While this inevitably produces many points that are in the ocean, these firm 0.0 ("No") data points are necessary for training the model. This numerical classification allows for a regression model to be fit to the data and thus establish probabilities for any testing point; however, a classification model was also ran for comparison.

For any occurrences without a location in latitude/longitude, I removed these points. Additionally, I removed depth as a predictor for all terrestrial species (*Parthenium* weed and the lanternfly), as well as elevation as a predictor for all aquatic species. However, just with removing the latitude and longitude, I found that my occurrences across all species ranged from 200 to just over 2000 per species, which can be suitable for a

machine learning model. However, not all points were able to successfully extract climactic variables, and if we had removed these instances, there would be insufficient data for a model. As such, I extrapolated random climate data based on the existing values.

Exploratory Data Analysis

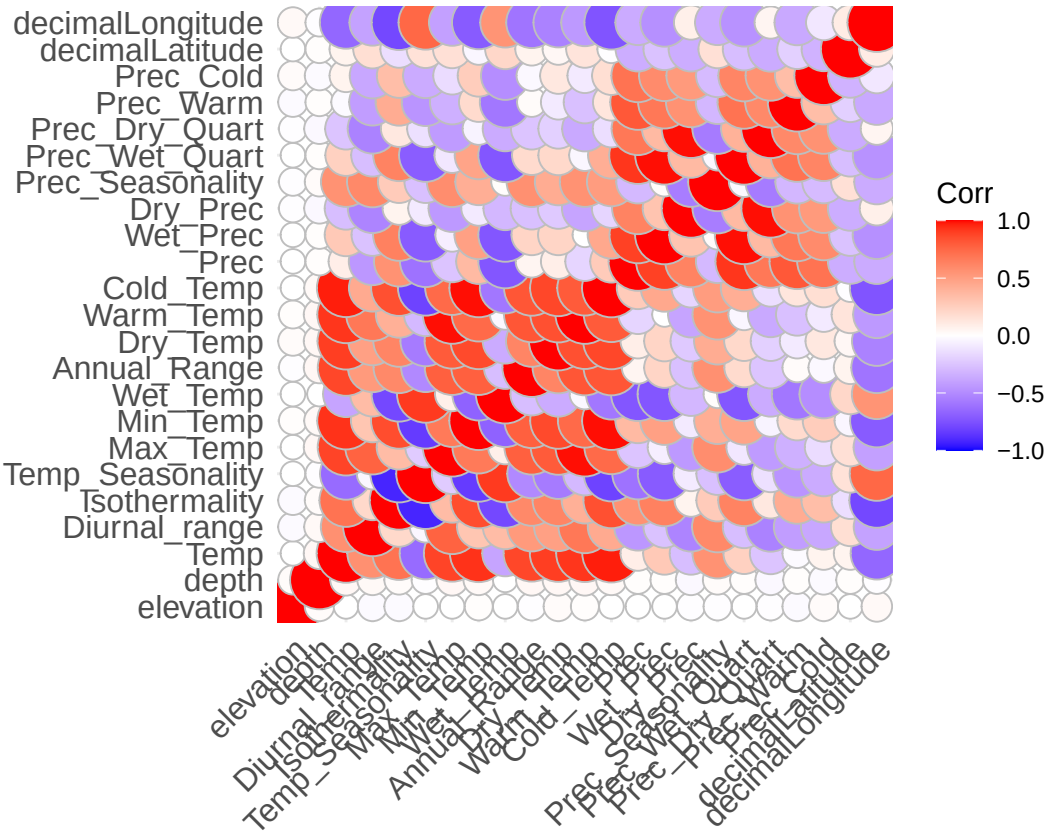
All EDA was conducted on actual occurrence data, not any data that was randomly generated of non-occurrences; any exploration of this would not contribute any particular insight.

First, I conducted a simple numerical EDA for all the data lumped together; for this, the processed data was read into Golang, and various functions I had written then outputted text files of summary statistics for the data, which are displayed in tabular form below.

Table 2: Numerical Summary Climatic Variables

Predictor	Mean	Variance	Standard Deviation	Min	25%	Median	75%	Max
Annual_Range	208.697498	8.736692e+03	93.47027	-243.00000	162.500000	241.000000	262.00000	375.00000
Cold_Temp	145.816528	1.665344e+04	129.04820	-375.00000	107.000000	190.000000	241.00000	280.00000
Diurnal_range	120.067475	1.010829e+03	31.79354	36.00000	95.000000	122.000000	147.00000	187.00000
Dry_Prec	17.603487	1.001063e+03	31.63959	0.00000	0.000000	2.000000	24.00000	309.00000
Dry_Temp	190.442760	1.418459e+04	119.09906	-375.00000	170.000000	235.000000	262.00000	359.00000
Isothermality	53.837756	3.439858e+02	18.54685	11.00000	39.000000	56.000000	69.00000	92.00000
Max_Temp	323.009856	7.971457e+03	89.28302	-53.00000	294.000000	328.000000	386.50000	484.00000
Min_Temp	77.260804	1.500569e+04	122.49770	-416.00000	32.000000	111.000000	166.00000	233.00000
Prec	824.194845	5.176085e+05	719.45013	0.00000	226.500000	679.000000	1205.000004636.00000	
Prec_Cold	168.635330	6.437697e+04	253.72617	0.00000	4.000000	48.000000	213.50000	2143.00000
Prec_Dry_Quar	64.914329	1.150539e+04	107.26319	0.00000	1.000000	13.000000	92.50000	1050.00000
Prec_Seasonality	73.211524	1.687398e+03	41.07794	0.00000	38.000000	70.000000	101.00000	261.00000
Prec_Warm	193.654283	3.649798e+04	191.04445	0.00000	32.000000	155.000000	288.00000	1276.00000
Prec_Wet_Quar	375.065201	9.329547e+04	305.44308	0.00000	118.500000	305.000000	604.50000	2143.00000
Temp	199.709629	1.012197e+04	100.60802	-224.00000	174.000000	238.000000	265.00000	314.00000
Temp_Seasonality	1116.811221	1.020260e+07	3194.15038	185.00000	1267.000003222.000006524.5000014779.00000			
Warm_Temp	250.694466	6.227936e+03	78.91727	-87.00000	224.000000	262.000000	303.50000	380.00000
Wet_Prec	144.014405	1.328855e+04	115.27599	0.00000	47.000000	120.000000	226.00000	886.00000
Wet_Temp	245.749052	8.338615e+03	91.31601	70.00000	171.000000	244.000000	313.50000	497.00000
decimalLatitude	17.713188	1.359571e+03	36.87237	-88.22555	- 6.963816	15.470162	54.09863	88.16152
decimalLongitude	-1.744985	2.741900e+03	52.36315	-178.69411	- 17.020289	- 0.033817	11.89516	174.98670
depth	69.789008	2.550182e+04	159.69288	0.00000	17.000000	38.900000	68.00000	2200.00000
elevation	254.487395	2.139649e+04	146.27541	0.00000	125.000000	259.000000	381.00000	500.00000

Secondly, I ran a correlation matrix to identify variables that are highly correlated with each other; this again involved usage of Golang to output numerical data, while a visualization was created in R.



We can see that a lot of the variables are highly correlated with each other. This makes sense; after all, weather patterns are usually dependent on geographical features as well as each other. Additionally, many of the variables are variations on temperature and precipitation. As such, it may make for an over-fitted model.

The point of all of this EDA is two-fold. Firstly, it allows users to understand any correlations that exist within the data that could otherwise lead to an over-fitted model. Based on this, we can establish a threshold for correlation for which, if a pair of variables has a correlation above this threshold, then the preprocessing recipe can remove it so as to not get an over-fitted model. Secondly, looking at the data numerically allows for a quick and easy way to understand the data before it is fit into the model and potentially catch any outliers or inconsistencies within the data. However, since the data is from reputable sources, we would expect it to be fairly clean.

Model Development

For model development, I used built-in functions in the `tidymodels` and the `biomod2` R packages. For `tidymodels`, this involved creating a data processing recipe. This mostly removed all non-predictor variables and also used bagged trees to impute missing data, a method that can somewhat reliably generate new data points based on the ones that already exist.

Following this, I simply split the data for each species into a training and testing set. I then fit a generalized linear model (GLM), random forest, and boosted forest as regression models to the data for all applicable species. A brief explanation of these is given as follows.

Generalized Linear Model

The GLM, in short, is an algorithm that builds off of ordinary linear regression. With classical linear regression, our goal is to minimize the sum of squared errors (i.e., the total distance between the fitted

value and the actual value for all points). While ordinary linear regression forms models that are simple to understand, it is a fairly "rigid" test, at least from a statistical point of view.

For ordinary linear regression, statisticians assume that the relationship between the two variables x and y is given as follows:

$$y = \overrightarrow{\beta} x + \epsilon,$$

where $\overrightarrow{\beta}$ represents a vector of predictor values (slopes) corresponding to the vector of predictor variables x, and ϵ represents the error between the model and the actual values. These errors are normally distributed in classical linear regression, and their expected value is zero. However, the GLM allows for these errors to come from a variety of distributions.

Thus, for the purposes of understanding the GLM for this project, the y value is either a 1 or a 0 corresponding to if a species is present or absent in a given location, and the GLM will simply fit a line to the data to try and minimize the error between itself and the given points. However, for the purposes of this project, the main things to know is that any model produced can be explained with any returned coefficients (though I will not do so in this project), and that this model is computationally very easy to conduct, but it lacks in terms of accurate results.

Random Forest

To understand how Random Forests work, we first need to briefly discuss decision trees. Decision trees are a simple machine learning model that can be used for classification or regression; they form binary trees where each node corresponds to a boolean expression, and a specific child path from that node is taken dependent on the true/false value of that expression. Once a leaf is reached, the decision tree will either tell a specific value (if a regression model is used) or a class (if a classification model is used).

Decision trees are easy to visualize and explain; additionally, they are not computationally difficult to fit to data. However, they have the innate problem of overfitting a training set; this is due to their tendency, like the GLM, to minimize error with the training set, which can create a tree that will not work on data outside the training set.

Random forests, in short, remedy this problem by taking many of these decision trees and aggregating their decisions together (hence "forest"). These decision trees are decided more randomly than classical decision tree algorithms (hence "random"), and the final result is the mean of the predicted values for all the trees (in regression) or the most commonly predicted class (if classification).

For the purposes of this project, the main things to know is that this model is moderately computationally expensive and not particularly explainable, but the results are of good quality and avoid over- and under-fitting.

XG Boosted Forest

The XG Boosted Forest is a high-performing model that is commonly used in a wide variety of applications, including both classification and regression. On a very high level, it starts by initializing all fitted values with the most common class (in classification) or the mean response value (in regression), though some algorithms will also start with a single tree fitted to the data.

Following this, the algorithm then computes the value of the gradient of the errors at this point based on the current model. A second tree is then calculated based on the values of the gradient, aiming to further reduce the total error. This tree is then combined with the first to form a new model, and this process is repeated until we either go below an error threshold or we reach a specified maximum amount of trees.

For the purposes of this project, all that is necessary to know is that this model is very computationally intensive and can easily crash a computer that does not have sufficient memory; additionally, the model is not very explainable. However, the payoff here is that the results are generally very good and speak for themselves.

Model Results

The metrics for replicating the *Parthenium* ML model for each type are given as follows:

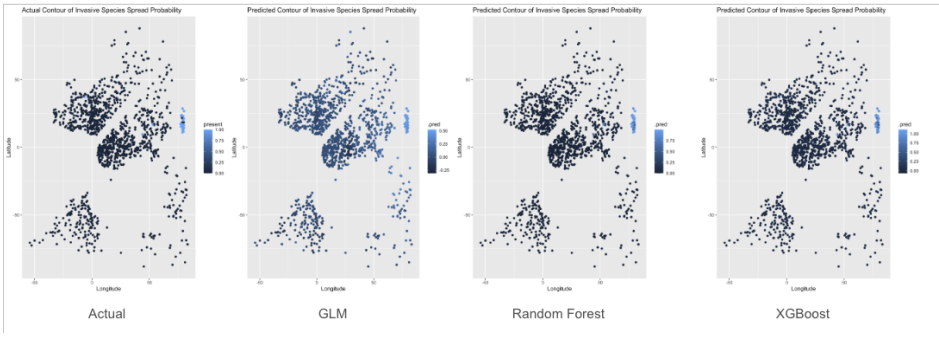
```
## # A tibble: 6 x 3
##   .metric .estimate method
##   <chr>       <dbl> <chr>
## 1 rmse       0.150  Generalized Linear Model
## 2 rsq        0.436  Generalized Linear Model
## 3 rmse       0.0614 Random Forest
## 4 rsq        0.907  Random Forest
## 5 rmse       0.0483 XG Boosted Forest
## 6 rsq        0.942  XG Boosted Forest
```

Following this, I applied a classification model to the same data, but the capabilities of `tidymodels` do not allow for classification on a Generalized Linear Model, so I only used the random and boosted forest.

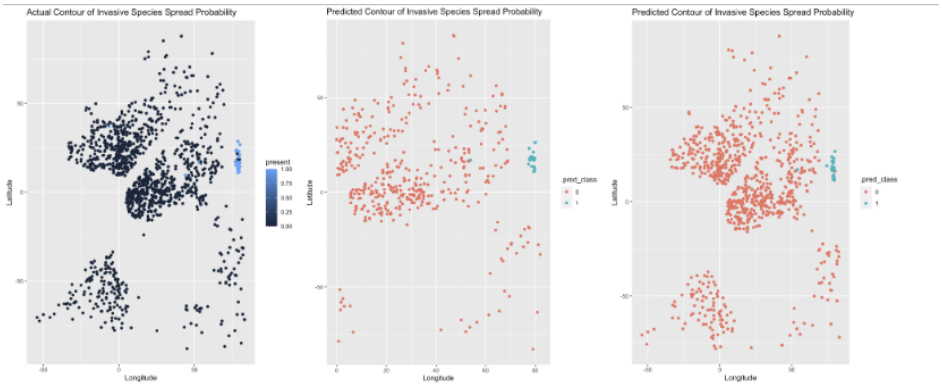
The metrics for classification are given as follows:

```
## # A tibble: 4 x 3
##   .metric  .estimate method
##   <chr>        <dbl> <chr>
## 1 accuracy  0.998    Random Forest
## 2 roc_auc   0.000127 Random Forest
## 3 accuracy  0.998    XG Boosted Forest
## 4 roc_auc   0        XG Boosted Forest
```

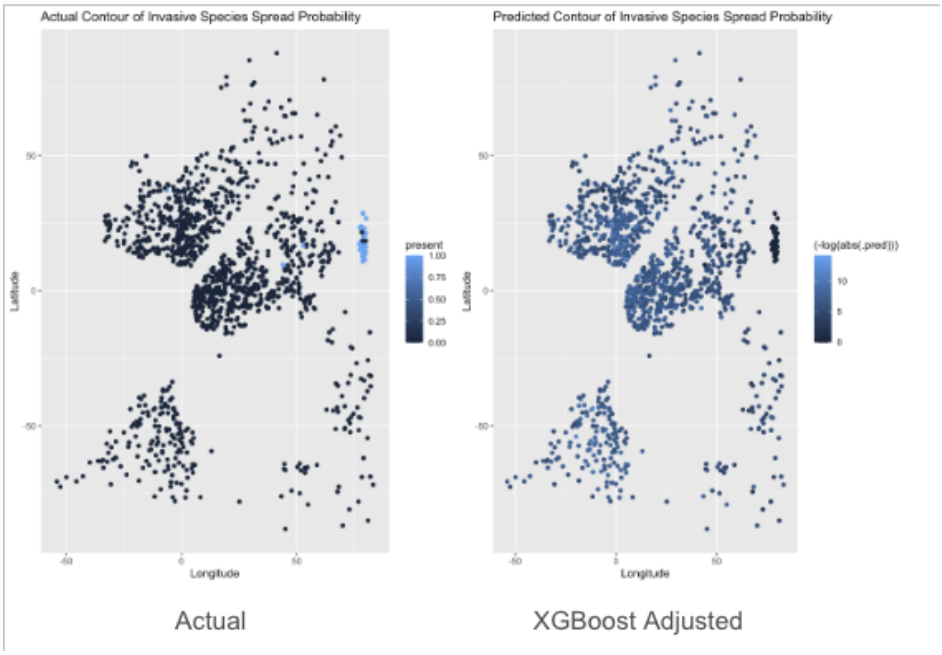
Following this, we would determine that the model is rather accurate, but there is the possibility of simply predicting all entries in the test set as being not present. As such, we need to make a plot to truly map the efficacy of these models. Since the boosted forest performed the best for both regression and classification, we will use that for plots. First, we show the regression plots, with contours mapped to the predicted values (i.e. probabilities).



Next we show the classification plots:



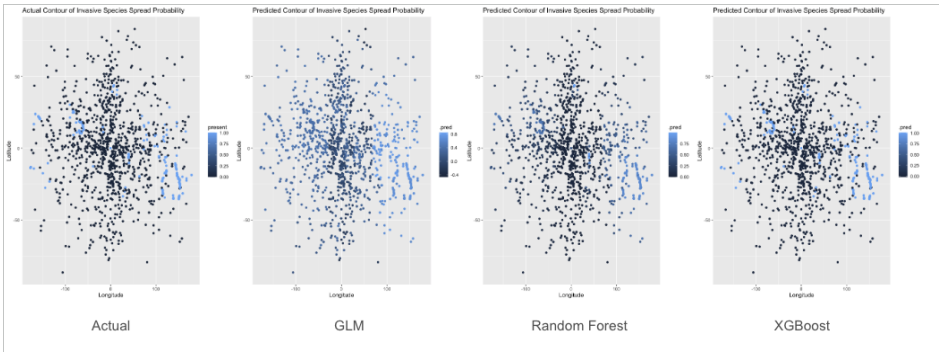
We can see that in general, there is a large portion of "no" points where we said the species was not present; however, there is clearly some accurate prediction with the regression model as we are able to have some band where there is accurate prediction of where the species are present; this is especially prevalent with the Random Forest and XGBoost models, which tend to be more accurate and predict around 0 and 1. To visualize probabilities more accurately with these models, we would need to alter the scales of the regression values, so we can try that here with the XGBoost plot.



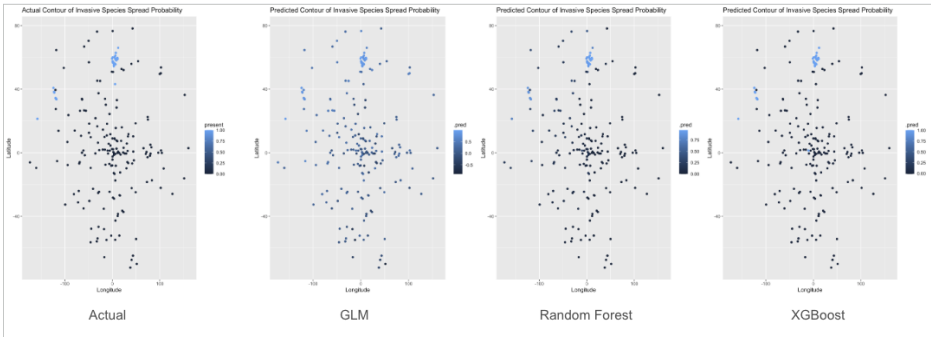
By taking the negative log of the absolute value of the regression value, we get larger values as the model predicts a species is less likely to be present in a certain point, and we can see more of a gradient present. Here, the gradient is such that a lighter point means an organism is less likely to occur at that location. Such a gradient normally exists with the GLM model as it is linear, so there is no need to adjust there. As such, we can conclude that while the GLM model is less accurate, it can be better in this scenario when modeling spread likelihoods.

We then tried all models with the remaining species. Note that images are avaliable in full resolution in the images folder provided alongside the code, including adjusted plots.

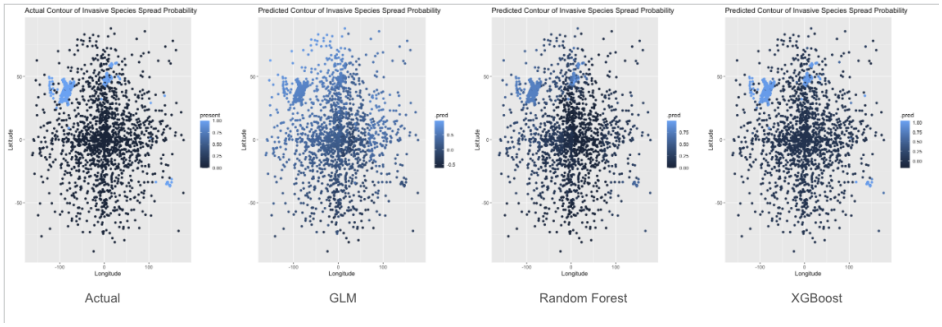
First, *Caulerpa taxifolia*:



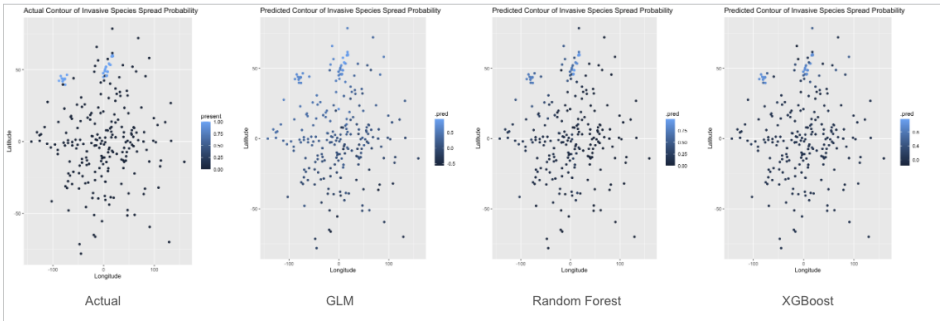
Then, *Ciona intestinalis*:



Then, *Craspedacusta sowerbii*:

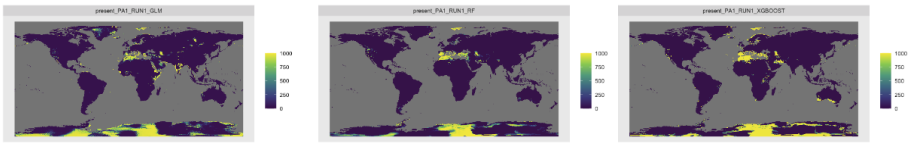


And finally, *Dreissena polymorpha*:

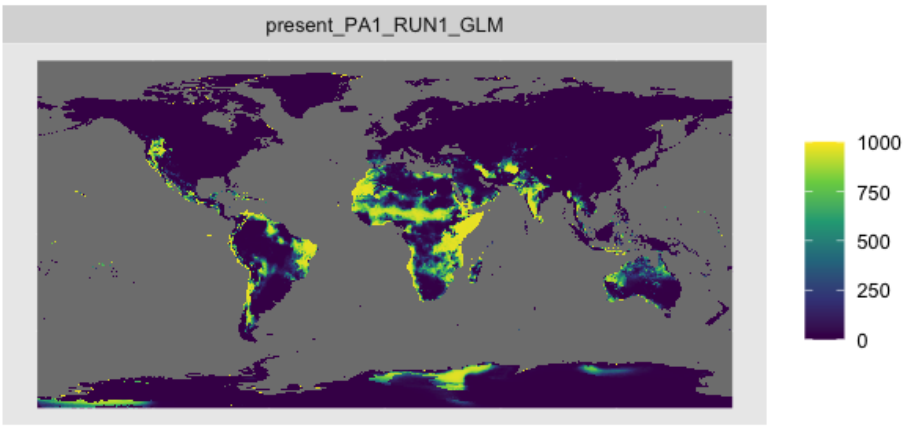


The models failed when ran on lanternfly occurrences; this is likely due to the fact that occurrences were filtered out due to no latitude/longitude. This forms an inherent difficulty in understanding occurrences as most people do not have the latitude/longitude handy in any given location.

In comparing the models to the models formed by `biomod2`, we see results similar to the following. Note that these models produce better maps, which is a feature of the package that I was unable to implement in time.



On this modeling, we can see somewhat similar results; though the `tidymodels` results do not align directly with the `biomod2` based models, there is still some similarity in that places tend to be predicted as having nothing or a high number. Additionally, these models predict many points in Antarctica as likely for spread, which is not accurate. Finally, I show a model for an aquatic species (again, all `biomod2` results are located in the supplementary images):



This model result in particular highlights the difficulty that `biomod2` has with occurrences of aquatic species; this is understandable given that it tends to work with climactic variables and other raster data which are more common for land. However, the `tidymodels` model forms a unique approach to this in highlighting

points in the ocean, which can help with further spread modeling for aquatic organisms.

Conclusion

Overall, the `tidymodels`-based machine learning models provided a fairly accurate projection of predictions for invasive species spread, especially for the *Parthenium* weed; while I was unable to exactly match the work done by Mainali et al., I found that putting my own spin on this complex topic to be rewarding nonetheless.

The `biomod2` and `raster` packages are relatively difficult to learn and implement, requiring many different file types and gathering data from many different places on the internet. While `tidymodels` is not easy in and of itself, it is more applicable to more R users and as such can be used to bring light to the complexities behind invasive species modeling and why it is so imperative to do what we can to mitigate their spread.

That being said, I would without question caution against a model such as the one I have created being used in real-world decision making; it could use much more refinement, especially in the acquisition and generation of data. In fact, if there is one main thing I learned from this project, it is that while data on the environment and species prevalence is easy to obtain, it is difficult to manipulate. Additionally, I would advise additional predictors being used in a more real-world focused model, especially for marine species (e.g., water conditions, boat traffic, etc).

For future work if I were to continue this project in the long-term, I would focus on three main aspects: field data, more predictors, and outreach.

Field data is by far the best way to evaluate these models, though it is labor- and time-intensive. Field studies to gather detailed information about specific points combined with expert domain knowledge on various species would give more accurate likelihoods of an organism appearing there than this model, despite what a model may say about its "goodness" via its metrics. In the `tidymodels` case, many of the points fed to the model were 0's and 1's, so it predicted around those values anyway and caused the metrics to seem good. Because we have to take any metric for these models with a grain of salt, the best method for evaluation is field data.

Secondly, more predictors. I chose various invasive aquatic species due to their prevalence and danger and was met, especially with `biomod2` usage; as mentioned previously, more predictors would make for more accurate models and would especially help in modeling the spread of aquatic (particularly marine) species. Predictors could include direction of water flow, light content, turbidity, boat traffic, and others; this would be hard to obtain but I hypothesize it would make for a more accurate model.

Finally, many people see these sorts of spread maps without understanding the work that goes on behind them and the huge necessity for quality data. As such, some outreach about the modeling process, including potentially touring datasets and making models of their own, could convince the public to provide more quality data of occurrences when they see an invasive species. Additionally, some people do not even know certain species are invasives, so outreach would also help in the main thing this modeling work aims to do: prevent spread.

References

Mainali, K. P., Warren, D. L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., ... & Parmesan, C. (2015). Projecting future expansion of invasive species: comparing and improving methodologies for species distribution modeling. Global change biology, 21(12), 4464-4480.

GBIF.org (13 October 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.3f7tcw

GBIF.org (29 October 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.q3tqpy

GBIF.org (29 October 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.wavs55

GBIF.org (29 October 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.nbrags

GBIF.org (29 October 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.56gqnm

GBIF.org (29 October 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.d85sdb

Fick, S.E. and R.J. Hijmans, 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37 (12): 4302-4315.

Huug van den Dool, Jin Huang and Yun Fan. Performance and Analysis of the constructed analogue method applied to US soil moisture applied over 1981-2001. HJ. of Geophysical Research, vol. 108, 2003, p 1-16

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., and Townshend, J.R.G., 2013, High-Resolution Global Maps of 21st-Century Forest Cover Change: Science, v. 342, no. 6160, p. 850-853.