Using Machine Learning to Predict Narragansett Bay Fish Trawls Based on Phytoplankton Community Composition

Cassy DeBlois
University of Rhode Island

Abstract

Fish community populations are affected by many different factors and change from year to year, especially with habitat loss and migrations due to global warming. However, no matter where fish migrate, they still need prey that rely on the phytoplankton in the water column to survive. The goal of this project was to see if phytoplankton community composition could be used to make predictive models about fish in Narragansett Bay. While almost all of the models seemed successful, with 90% accuracy or more, the models should not be trusted because of the low variability and the low number of replicates in the data.

Introduction

- Over time, plankton communities are changing due to global warming making it more difficult for specialist species to thrive and increasing harmful algal blooms (Sarker et al 2020, Gobler et al. 2020)
- Fish communities are also changing over time with global temperature increases impacting sensitive species and causing habitat shifts (Punzón et al. 2016)

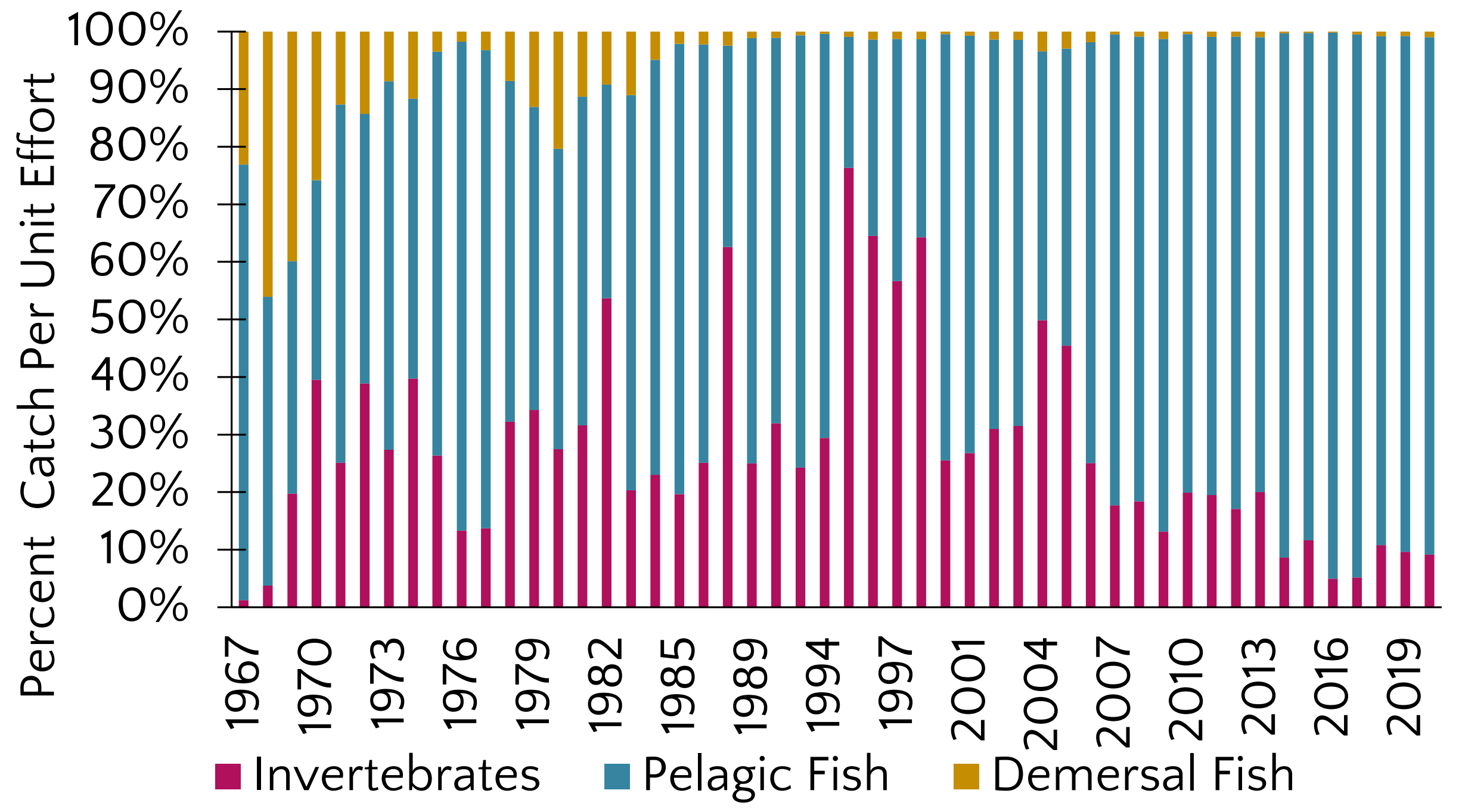

Figure 1. Percent catch per unit effort of fish trawl data over time. Even in three broad categories, major community composition changes can be seen.

- On top of specific temperature needs, fish also rely on their prey and their food web to survive which starts with phytoplankton meaning that changes in the plankton communities may lead to cascade effects on fish populations
- **The purpose of this project is to use phytoplankton cell counts to try to predict fish trawl data with machine learning models in python**

Methodology

The following data sets were used in this study:

- The Narragansett Bay Long-Term Plankton Time Series from the University of Rhode Island
- The Fish Trawl Survey from the University of Rhode Island – specifically the data from Fox Point as it is the most similar location to the above survey

Combined, I had **50 years** of data with **90 plankton species** and **25 fish species**

In excel, the data was cleaned and organized so that both sets of data could be integrated into one table. Years with null data points were removed from the data set as they cannot be processed by the machine learning algorithms used.

Then, the data was imported into python code that could further transform the data into a more useable form and could perform the machine learning algorithms.

Limitations

This project is meant to be an exploratory look at machine learning on complex community compositions and is by no means conclusive. There are many simplifications in the data to allow the machine learning to run on the available equipment and therefore there are many limitations in this study including:

- Several years and plankton species have missing data that were not included in this study because the machine learning algorithms cannot handle null numbers
  - There are also many plankton species taken out of the data set because the species from the 1959-1997 data set does not match the species from the 1999-2024 data set
- The phytoplankton cell counts are averaged over the course of the year to match the fish trawl data which does not account for seasonality in either database
- The continuous data was transformed into discrete bins to make predictions easier to run, but this also means the model is likely over-accurate since it has a much higher chance of being correct
- 50 years is incredible for a long-term data series! However, it is very little for a machine-learning model because the data has to be split into the training set and the testing set, leaving few replicates for each

Results

I tested several types of machine learning algorithms, but settled on using a decision tree classifier because they can more easily be visualized and then manually used to predict each fish population. They can be used similarly to a taxonomic key.

Can machine learning models be made from cell counts?

Yes! The machine learning algorithm was successful on 19 of the 25 fish species. Due to a lack of replicates and variation in those replicates several species, like with butterfish and cancer crabs.

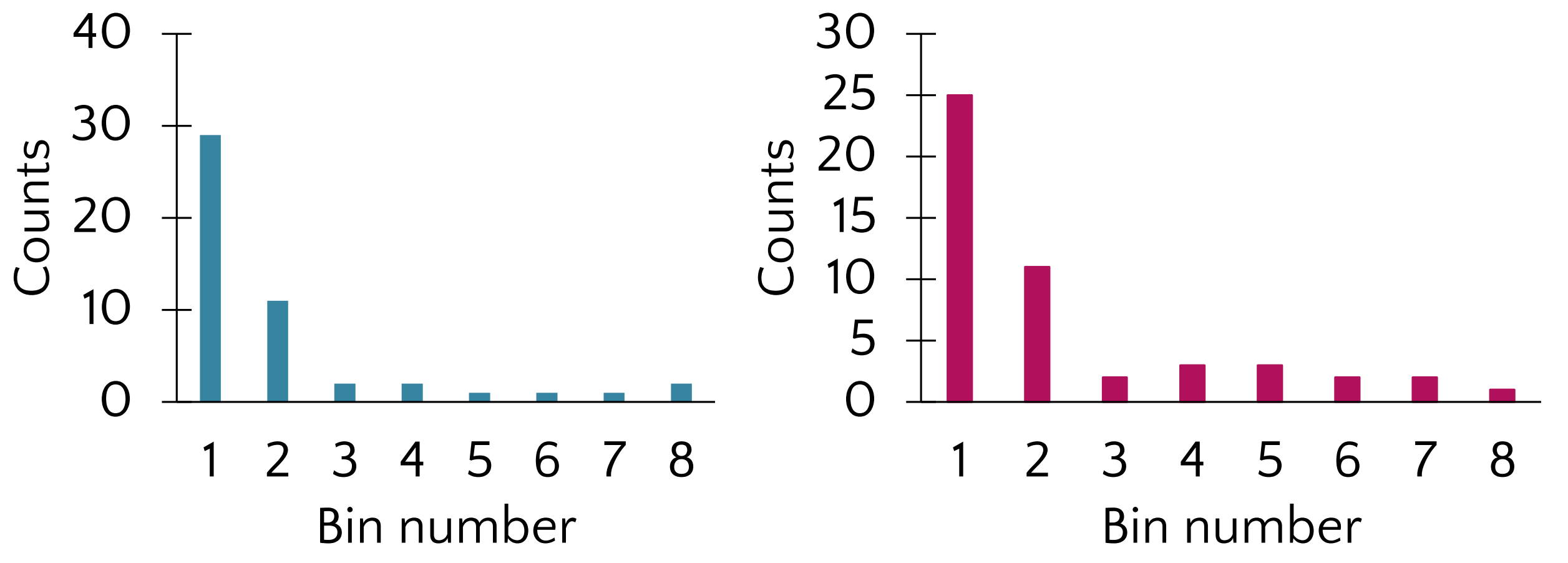

Figure 2 Histogram for the binned data for butterfish (left) and cancer crabs (right). Other than the first two bins, the counts for the rest of the bins are too small (<5), along with being split in half for training,

How successful were the models?

Table 1. Predictive accuracy of the decision classifier tree for each fish, all of them being 90% accurate or more.

Species	Alosa spp.	Atlantic herring	Bluefish	Cunner	Fourspot flounder	Horseshoe crab	Lady crab	Little skate	Longhorned sculpin
Accuracy	100%	95%	100%	100%	100%	100%	95%	100%	100%

Species	Northern searobin	Red hake	Sea star	Silver hake	Spider crab	Striped searobin	Summer flounder	Tautog	Weakfish	Window-pane
Accuracy	100%	100%	95%	100%	95%	100%	100%	100%	100%	90%

Should the models be used to predict future fish trawls?

No! Although the models look incredibly accurate given 90 different variables to determine a single fish count, accuracy is not the only part of the model that should be looked at because there is a particular reason that the models were so incredibly successful. Upon further analysis, the reason that the models are 100% accurate, like with Weakfish, is because there is very low year to year variability in the data. All of the data points fell within the same bin, meaning that as long as the model guessed that bin, it would always be correct.

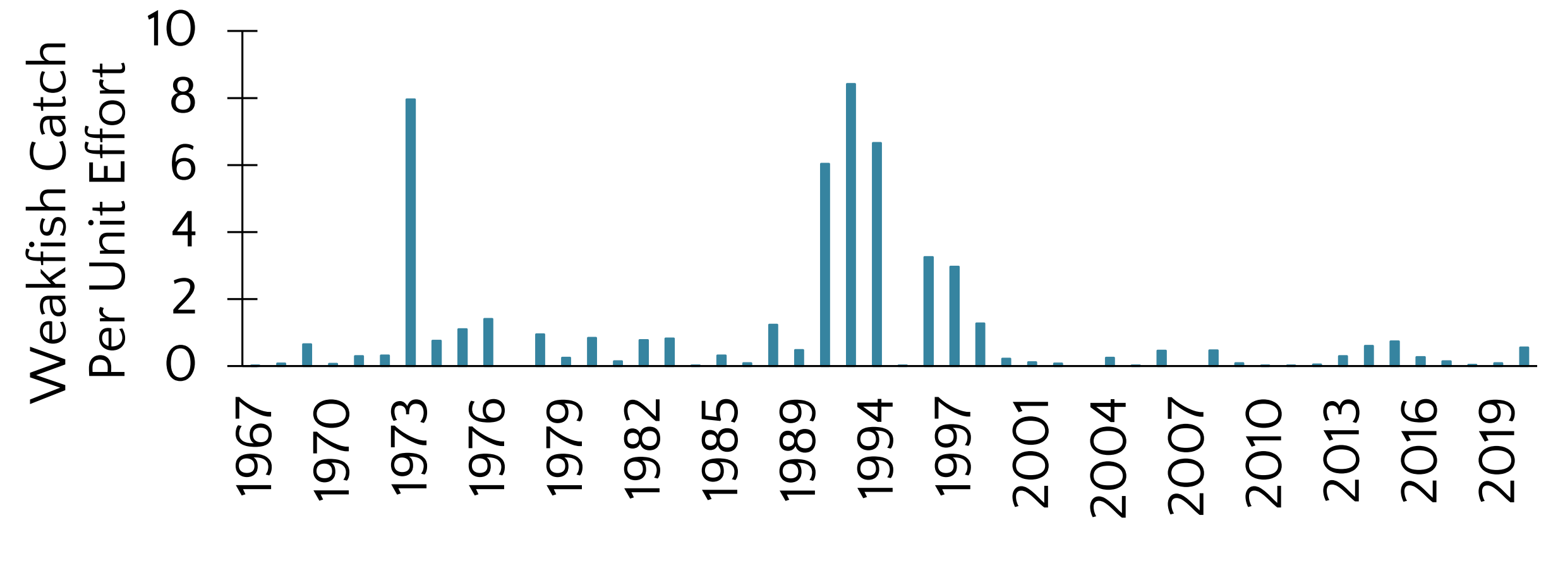

Figure 3. Yearly weakfish catch per unit effort. Note that the y axis only goes up to 10, and only four years are above 4 which is too few replicates for the model to use.

Conclusion

Lessons learned

- These models should not be used to predict fish trawl data from phytoplankton cell counts
- Machine learning models are a useful and powerful tool, but they need to be fully understood to be utilized properly
- Sample sizes from long term monitoring programs are too small for machine learning models to use, but they may just need more variability than typical fish trawls have to offer
- Plankton may just be a poor predictor for fish data! Although some of the correlations between the plankton and fish were >0.5 and <-0.5, there were strong correlations between plankton with other plankton species and fish with other fish species
  - For example with plankton, *Acinocyclus* has a correlation of 0.99 with *Amphidinium*
  - For example with fish, *Alosa* has a correlations of 0.95 with blue fish

Changes for if this project is attempted again

- More replicates
  - Yearly averages of the data severely reduce the number of replicates that can be used for training and testing the models along with lowering variability. In the future, the data could be separated into a shorter time frame, like each catch in the summer for example, to acquire more data points
- More variability
  - One idea for increasing variability could be to look at a more direct connection in the food web like phytoplankton and zooplankton counts since zooplankton will vary more in each catch than fish catches for the year

Acknowledgements

For information about me, my code, and my references, scan the QR code or email me at cdeblois_7355@email.ric.edu