Distribution of Six Aquatic Species in the Southern Gulf of Saint Lawrence, Canada

Statistics for Data Science

David Fishman[1]

[1]davjfish@gmail.com

April 2024

Contents

1 Objectives

Canada oceans provide and important source of livelihood to Canadian as well as fulfill important ecological services. The Canadian Government has a mandate to ensure the Oceans remain healthy and economically and ecologically viable. The Department of Fisheries and Oceans (DFO) is tasked with conducting regular stock assessments surveys. These survey can be carried out on either Canadian Coast Guard Science vessels or contracted through privately owned fishing vessels. The stock assessment surveys conduct fishing and oceanographic activity as a means to monitor fish populations and produce population indices. Many of the data from stock assessment surveys have been made available via the Open Government Portal [2].

1.1 Goal of the analysis

The importance of having robust models for predicting the distribution of aquatic organisms cannot be overstated. From a resource extraction point of view, modelled distributions can help direct fishing efforts; resulting in more efficient fisheries. From an environmental conservation point of view, modelled distributions can help identify which geographic areas should be targets for conservation efforts. The goal of this analysis is to build models that can be used to predict the probability of occurrences of species of interest in the Southern Gulf of St. Lawrence (sGSL). Specifically, I employ a logistic regression approach, using latitude, longitude and water depth (i.e., elevation) to predict the probability of occurrence of the following six species:

- American plaice (*Hippoglossoides platessoides*)
- Atlantic cod (*Gadus morhua*)
- Atlantic herring (*Clupea harengus*)
- Redfish unidentified (*Sebastes sp.*)
- American lobster (*Homarus americanus*)
- Snow crab (*Chionoecetes opilio*)

1.2 Rationale behind the analysis

Spatial autocorrelation, i.e., when observations in space that are closer together are more correlated, is a common occurrence in the natural world [6].

Accordingly, I would expect that using geographic coordinates to predict the distribution of occurrences would useful.

A basic tentant of ecology is that different species acquire different specializations and thus inhabit different ecological niches and strategies. In aquatic ecosystems, the physio-chemical and ecological environments change substantially with water depth. For example, certain species will be physiologically adapted to cope with the colder temperature, higher pressure and salinity levels that occur at low elevations. However, within the scope of a given ecosystem (e.g., within the sGSL) certain species will specialize more than others; i.e., specialist verses generalists. Accordingly, I would expect that, at least for certain species using a measure of water depth at a given location would be useful in predicting the likelihood of occurrence.

2 Data Preparation

2.1 What was your data source?

Three datasets were used in this analysis.

The first, as noted above, was the Government of Canada's Open Government Portal [2]. The name of the dataset is NAFO Division 4T groundfish research vessel trawl survey (September Survey) dataset [3]. The dataset contained information ecological information (i.e., species caught, specimen counts and specimen weights), fishing information (i.e., gear type used, fishing vessel) and spatial information (i.e., latitude and longitude). For mapping purposes, I also used a geospatial file of the Canada Provincial Boundaries from the same catalogue [4].

The dataset did not contain and elevation data and therefore this had to be acquired elsewhere. Luckily, there is an excellent website and web app called the General bathymetric Chart of the Oceans (GEBCO), which is maintained by the international community [5]. The web app provides an interface for downloading world ocean's bathymetric data.

2.2 How good was the data quality?

The quality of the fishing data was excellent. There were no missing values and the column data types were respected. For example, the columns that you would expect to be *float* type, e.g., lat/lon values, were indeed. The dataset was accompanied by a data dictionary, which contained clear explanation of the variables in both English and French. The only thing worth mentioning was that the coordinate reference system for the latitudes and longitudes were difficult to obtain. They are not displayed on the website but the metadata XML file did specify EPSG 4269.

2.3 What did you need to do to procure it?

The data from the Open Government Portal was downloaded using a web browser. The data was accessible via a static link. The format of the fishing data was a single comma separated values (CSV) document and the map was formatted as a GEOJSON file. The GEBCO elevation data was downloaded using their customizable web application. In this application, I was able to download the bathymetric data as a NetCDF file for only the area of interest. I was very impressed with this tool!

2.4 What tools or code did you need to use to prepare it for analysis?

2.4.1 Fishing Data

As noted above, the species / fishing data was very clean. I loaded the data from its original CSV into a pandas dataframe. I ensured all the data types made sense using the **DataFrame.info()** method. The dataset contained 166,694 rows of data; each one being a species observation at a particular place and time. When exploring the dataset, it became apparent that these for a given coordinate on a given date, there were many species observed. For the sake of efficiency, I parsed the original dataset into two separate dataframes: one containing the biological information (species observed, how many, total weight) and the other containing the site / fishing set details (time, date, lat/lon). The linkage between the two tables was made by a fishing set ID column named **site_id**. After this was done, I was dealing with a total of 7,257 fishing sets which are on display in Figure 1. Finally, I wanted there to be a boolean response column in the fishing set dataframe for each species of interest. To do this, I used the pandas *Series.apply* method on the *set_id* column. The apply function took the set ID and searched for a corresonding entry in the species dataframe. If one was found, the function return *True*, otherwise it returned *False*.

2.4.2 Mapping Data

The GEOJSON file was loaded using the geopandas python library into a geodataframe. The only data preparation needed for these data was to re-project the dataset to the same coordinate reference system as the fishing data. As noted above, the point data from fishing was presented in EPSG 4269. Without this step, the points and polygons cannot be represented on the same map.

2.4.3 Elevation Data

The elevation data was downloaded in NetCDF format and loaded directly into the python Xarray library [1]. When plotted using **Matplotlib**, you can clearly see the familiar contours of Atlantic Canada, the characteristic shallow waters of the sGSL and the deeper waters out in the channel of the St. Lawrence seaway (see Figure 2).

Using this two-dimensional array of elevation data, the next step was to extrapolate an elevation value for each fishing set in the fishing set dataframe. To do this, I followed an example from the **Xarray** docs. Specifically, I made use of the Data Array *sel* method. This method has an argument called *method* which specifies how to deal with inexact matches, i.e., when trying to determine elevation for a set of coordinates that are not in the array. I decided to use the *nearest* method, which returns value from the nearest coordinate. To add an elevation column to the dataframe, I used the pandas Dataframe.apply method in combination with the above. I then looked at a histogram of elevation value to: a) ensure that no single elevation value was greater than 0m (i.e. above sea level) and b) to get a sense of the distribution of elevation across the dataset (see Figure 3)

A screenshot of the first five rows of the final fishing set dataframe that was used in the subsequent phase of the analysis is presented in Figure 4

2.5 What challenges did you face?

Since geographic information systems is not my area of expertise, I found it challenging figuring out how to reproject a geographical data to another coordinate reference system. In the end, this was not too difficult, but it took time to figure out.

Another major challenge was learning about the Xarray python package [1]. This is a very impressive package, albeit quite complex. It took me a long time to figure out how to extract the data from a data array, manipulate it and create a new data array.

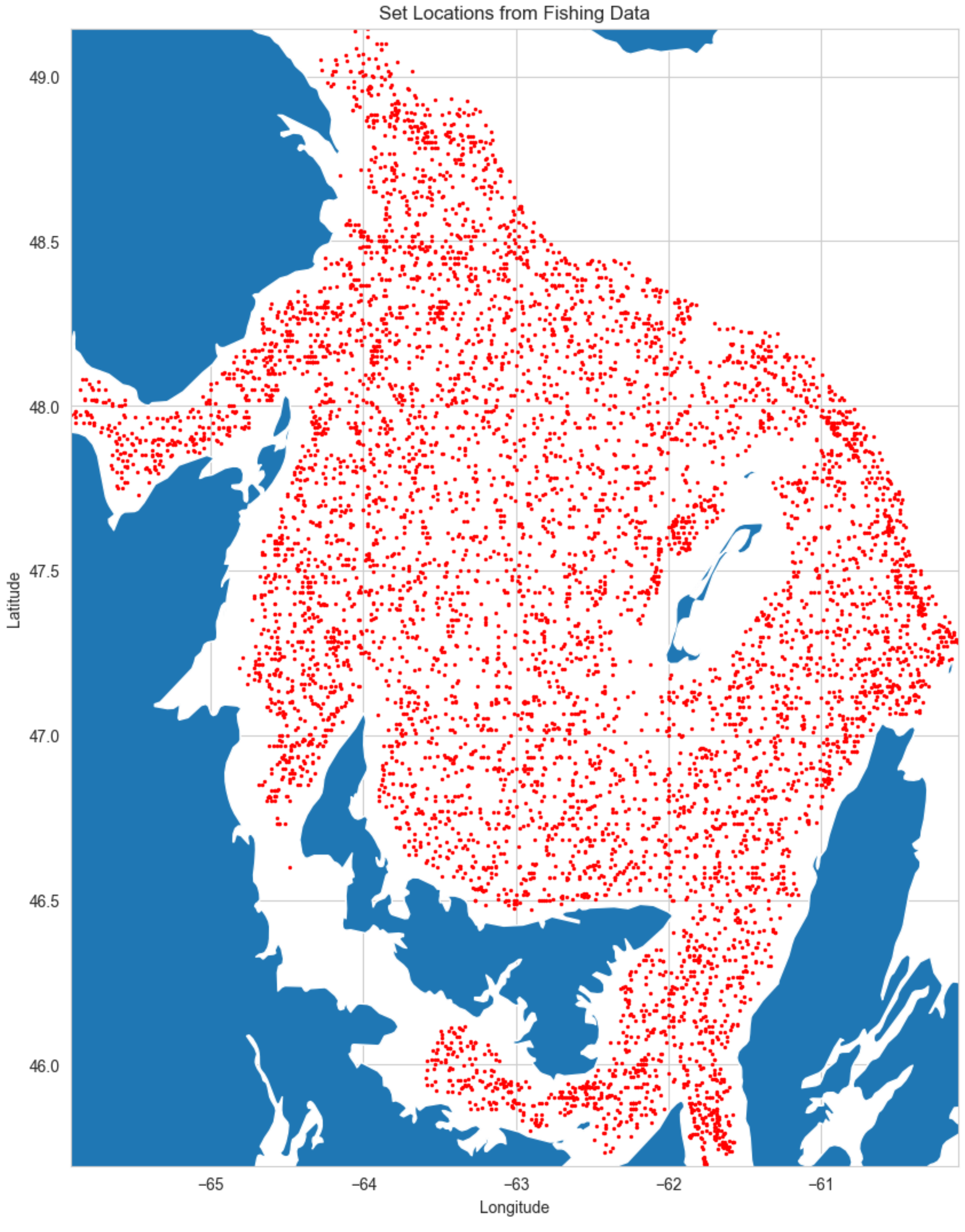


Figure 1: This figure displays the 7,257 fishing sets contained in this dataset. A fishing set is represented by a red point.

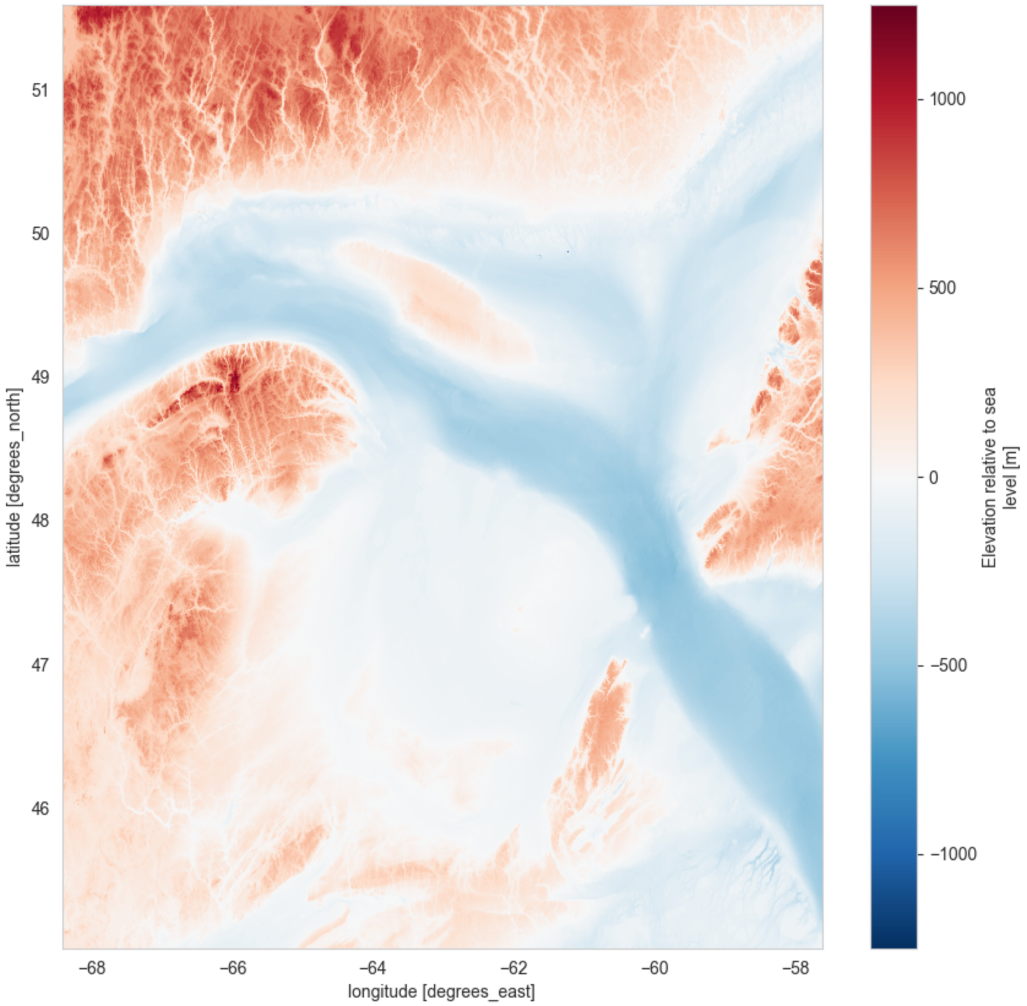


Figure 2: This figure shows the raw elevation data array. The values are color codes according to the elevation in meters

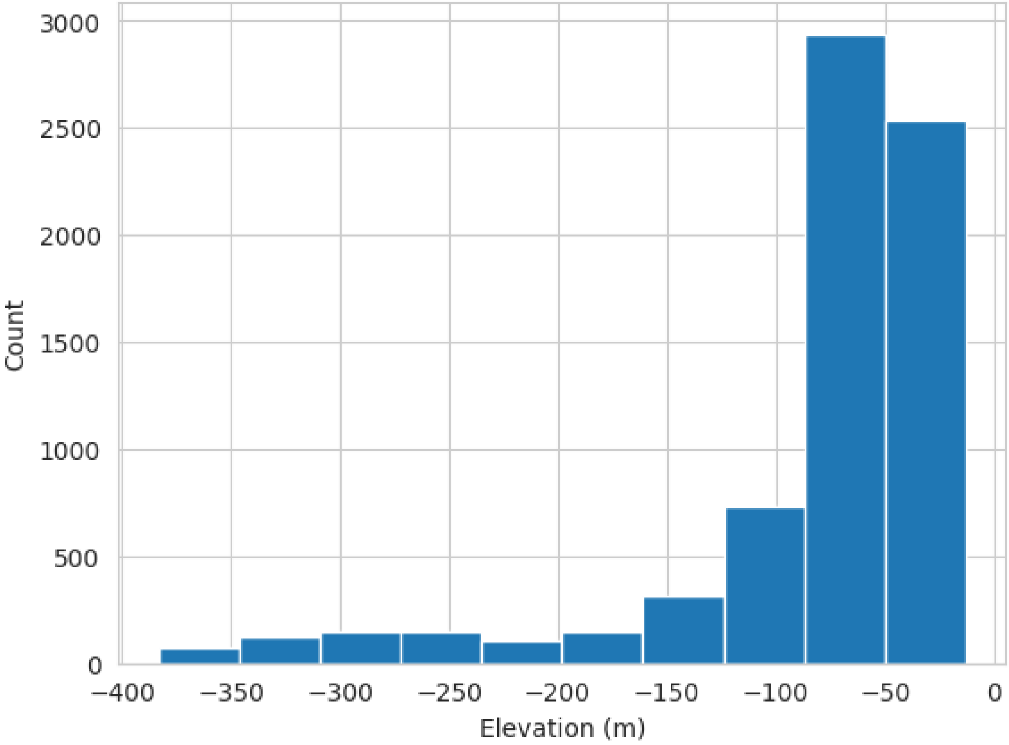
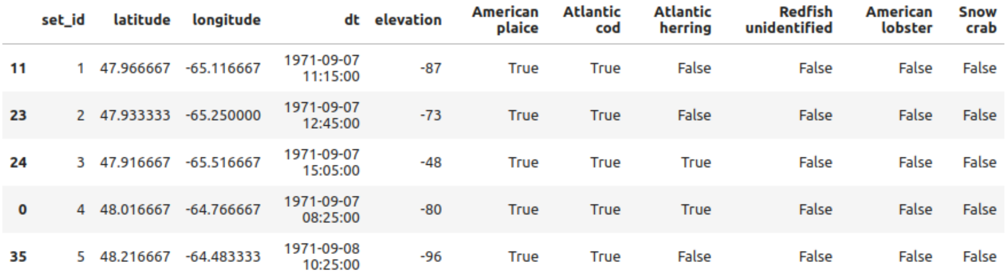


Figure 3: A frequency distribution of elevation in meters. Most of the sites have an elevation between -100m and 0m.



	set_id	latitude	longitude	dt	elevation	American plaice	Atlantic cod	Atlantic herring	Redfish unidentified	American lobster	Snow crab
11	1	47.966667	-65.116667	1971-09-07 11:15:00	-87	True	True	False	False	False	False
23	2	47.933333	-65.250000	1971-09-07 12:45:00	-73	True	True	False	False	False	False
24	3	47.916667	-65.516667	1971-09-07 15:05:00	-48	True	True	True	False	False	False
0	4	48.016667	-64.766667	1971-09-07 08:25:00	-80	True	True	True	False	False	False
35	5	48.216667	-64.483333	1971-09-08 10:25:00	-96	True	True	False	False	False	False

Figure 4: A screenshot of the first five rows of the fishing set dataframe, used in the susequent phase of the analysis.

3 Analysis

3.1 Logistic Regression Models

3.1.1 Methodology

The fishing dataset, as prepared in the previous section, was used to model the probability of observing each of the six species. A separate regression model was computed for each species, all taking on the following form:

$$Logit(pi) = \beta_0 + \beta_{lat}x_{lat,i} + \beta_{lng}x_{lng,i} + \beta_{dpt}x_{dpt,i}$$

- Where β_0 is the model intercept
- and $x_{lat,i}$ is the latitude of the i-th observation in the dataset, measured in decimal degrees
- and $x_{lng,i}$ is the longitude of the i-th observation in the dataset, measured in decimal degrees
- and $x_{dpt,i}$ is the elevation / depth of the i-th observation in the dataset, measured in meters
- and β_{lat}, β_{lng}, β_{dpt} are the model coefficients for latitude, longitude and depth, respectively
- and $Logit(pi)$ is the link function output for the i-th observation of the dataset.

The models were computed using the Python *statsmodels* package. A column of ones was added to the dataframe of predictor in order to represent the intercept. The resulting model for each species was then used to calculate the estimated probability of detection for each point in the dataset. This was achieved using the resulting model's *predict* method. Next, plots were generated for each species, displaying the relationship between the $Logit(p)$ function and the estimated probability of detection. prior to plotting, the rows of the data were sorted by the estimated probability in ascending order. In the same graphs, the observed presence-absence values were plotted in order to provide a sense of the distribution of the estimated probabilities across the true values. Boxplots displaying the same distribution of probably across the true presence-absence value were also generated.

Receiver operator curves (ROCs) were generated for each species-model. The ROC helps to understand the trade-offs in selecting for different discrimination thresholds in terms of obtaining true positives vs. false positives.

Obtaining more true positives means that you have to be willing to accept the probability of receiving false positives. For each ROC, the area under the curve (AUC) is calculated. As noted in our course module, a model with an AUC less than 0.5 performs worse than a random estimator. The larger the AUC, the better the model performance.

Finally, the main output of this analysis is a distribution map for each species' occurrence in the southern Gulf of St. Lawrence. For each coordinate in the data array, a corresponding estimate of probability of detection is calculated. The heat map provides a way to visualize the estimated probabilities over the landscape. In order to do this, I flattened the elevation data array into a pandas dataframe with the following columns: *const*, *latitude*, *longitude*, *elevation.* Then, once again using the model's **predict** method, I calculated the estimated probability of occurrence for each row. The resulting series of probability was then reshaped into the original ndarray shape of the elevation data array and inserted into a new data array containing the same coordinates as the original. These data arrays were then plotted using a heatmap theme color mapping. Finally, the true occurrences were overlain the heat map to provide a visual sense of model performance.

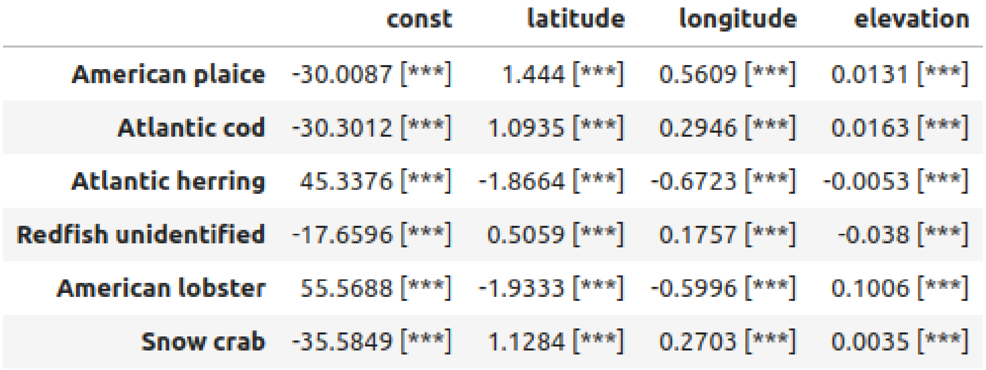
3.1.2 Results

The estimates coefficients for each model for the intercept, *latitude*, *longitude* and *depth* are presented in Figure 5. The estimated coefficients for all predictors were observed to be statistically significant at the $\alpha = 0.001$ level for each species and thus retained in the final model.

The resulting $Logit(p)$ vs. estimated probability plots for each model is presented in Figure 6. Visual inspection of the plots gives us a sense about how effective the resulting models will be at effectively

The distribution of estimated probabilities across the observed presence-absence values are presented in Figure 7.

The ROC plots for each species-model is presented in Figure 8. The shapes and corresponding AUCs are indicative of the performance of each model. All the computed AUCs are well-above 0.5, indicated that these models have at least some degree of usefulness. The two most powerful models produced were for Redfish and Lobster which had AUCs of 0.90 and 0.95, respectively. The least powerful model was for Snow Crab which had an AUC of 0.67. A table of computed discrimination thresholds for each model are presented in Figure 9. Finally, the resulting distribution map for each species in the sGSL is presented in Figure 10.



	const	latitude	longitude	elevation
American plaice	-30.0087 [***]	1.444 [***]	0.5609 [***]	0.0131 [***]
Atlantic cod	-30.3012 [***]	1.0935 [***]	0.2946 [***]	0.0163 [***]
Atlantic herring	45.3376 [***]	-1.8664 [***]	-0.6723 [***]	-0.0053 [***]
Redfish unidentified	-17.6596 [***]	0.5059 [***]	0.1757 [***]	-0.038 [***]
American lobster	55.5688 [***]	-1.9333 [***]	-0.5996 [***]	0.1006 [***]
Snow crab	-35.5849 [***]	1.1284 [***]	0.2703 [***]	0.0035 [***]

Figure 5: The estimated model coefficients for the intercept, latitude, longitude and elevation/depth for all six species. All coefficient in all models were observed to be significant at the $\alpha = 0.001$ level, as indicated by the three asterisks

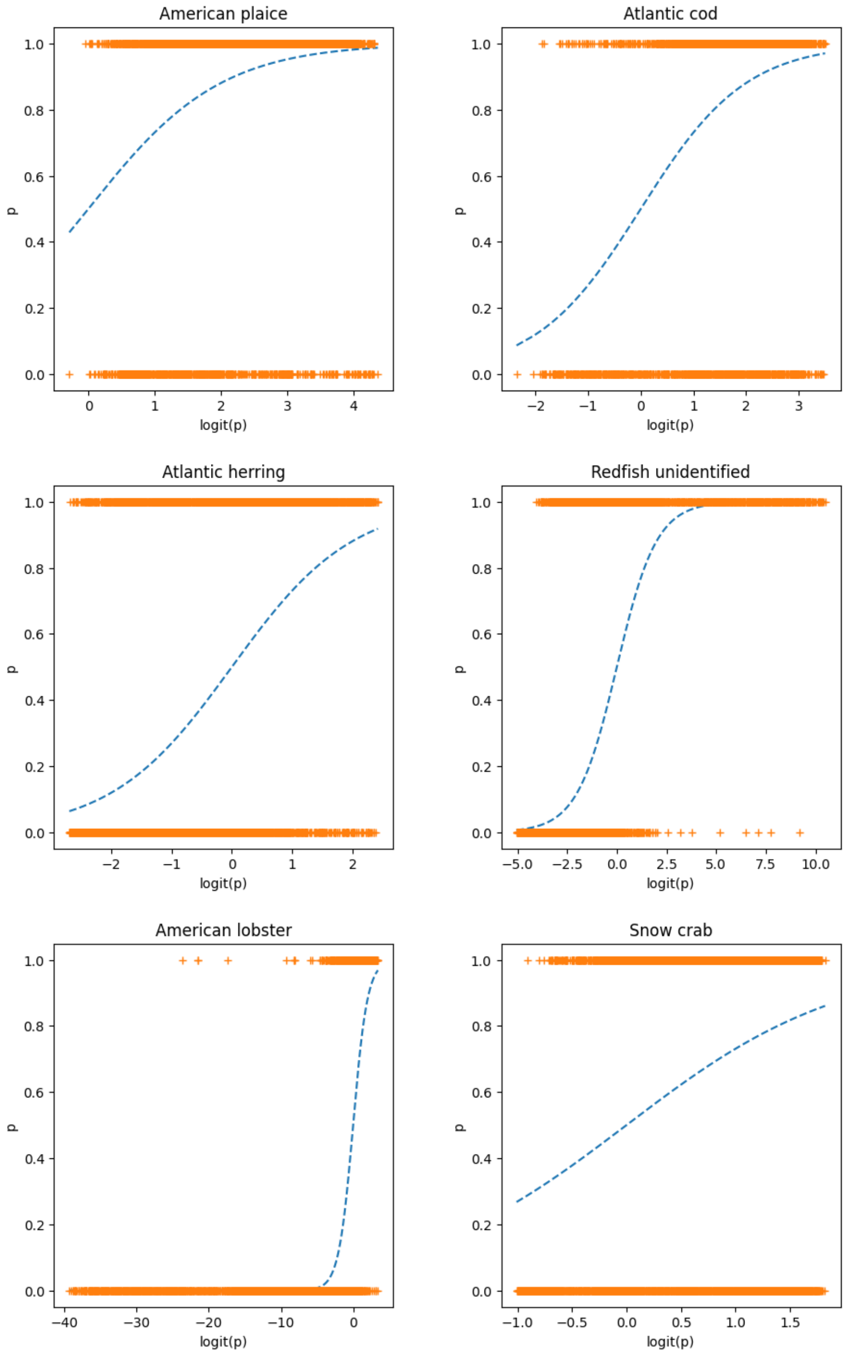


Figure 6: Plots displaying $Logit(p)$ vs. estimated probabilities for each resulting model. Prior to plotting, the dataset was sorted in ascending order of estimated probability. The observed presence-absence values are superimposed on the plot in order to provide a sense of the distribution of true values accross estimated probabilities.

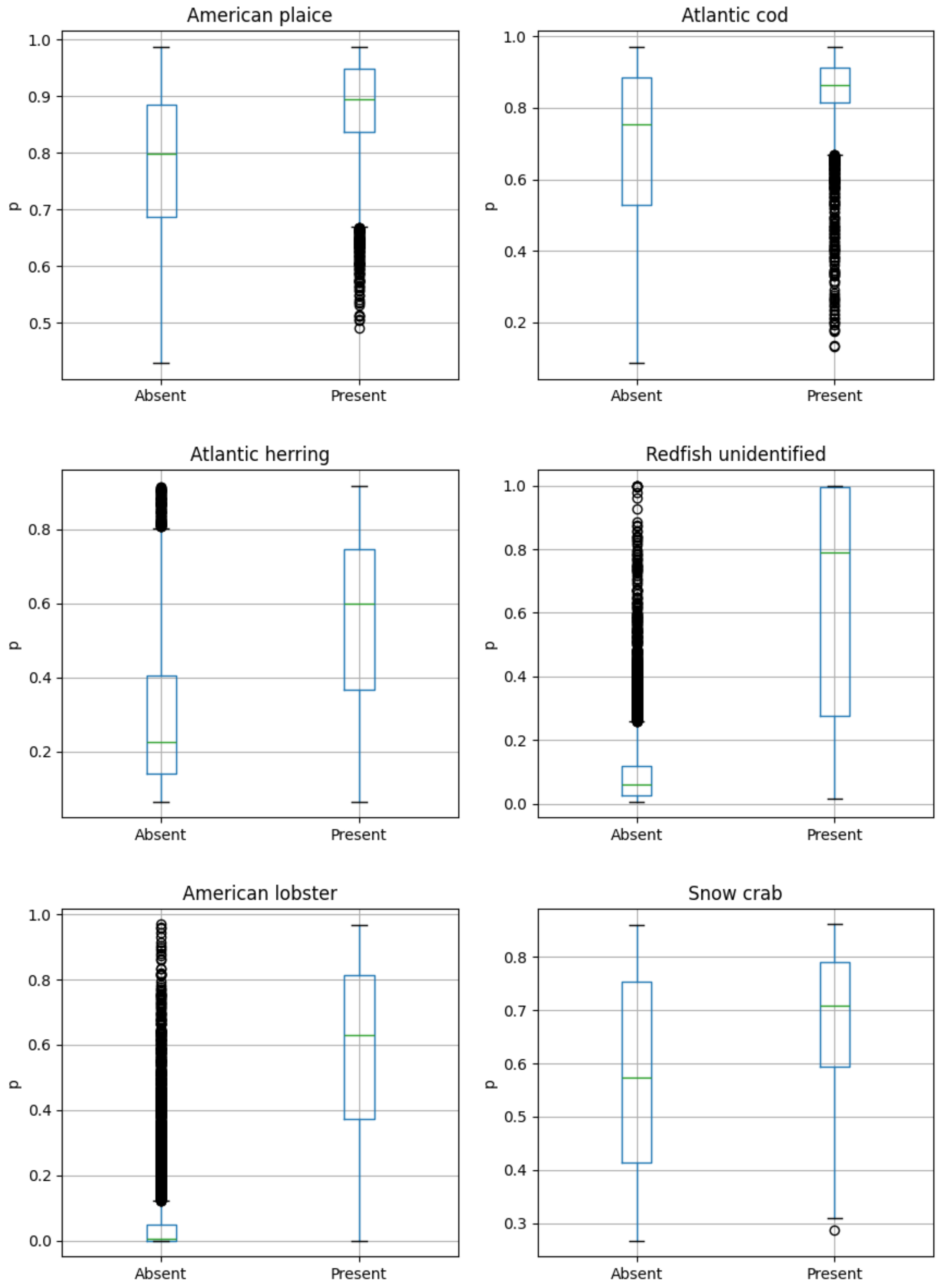


Figure 7: A boxplot for each species model that shows the distribution of estimated probabilities for presences vs. absences.

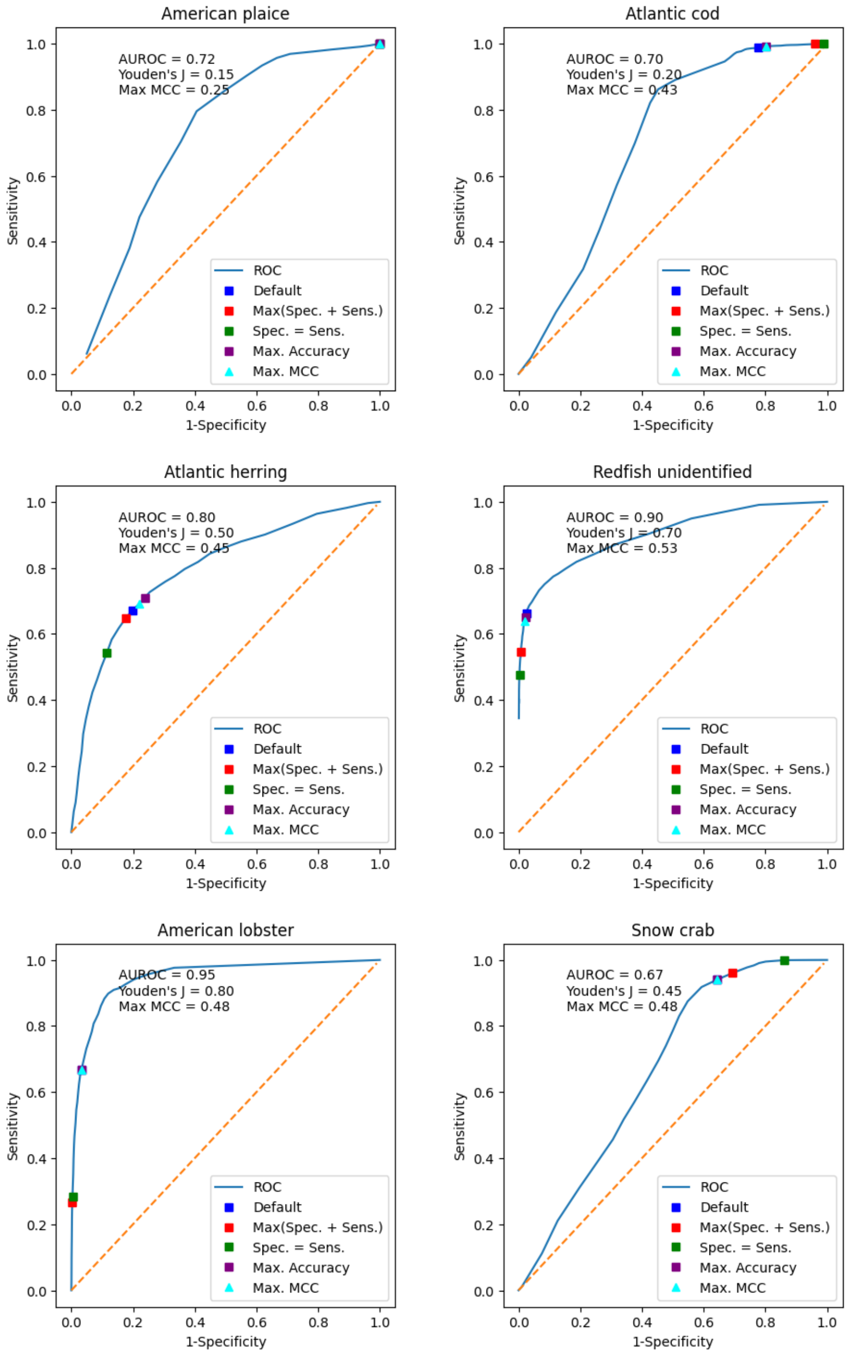
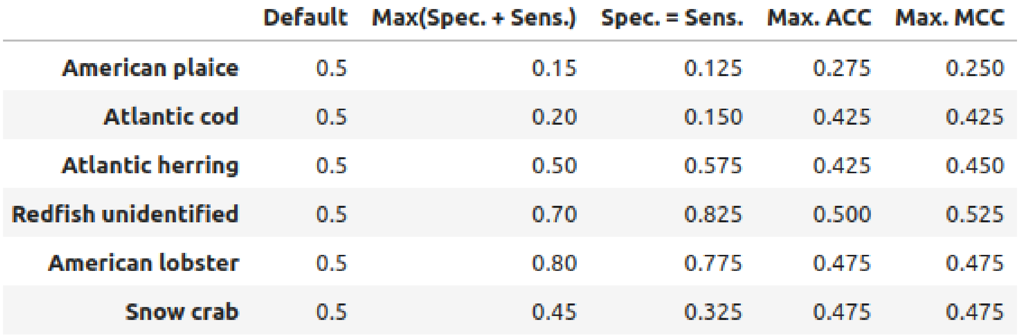
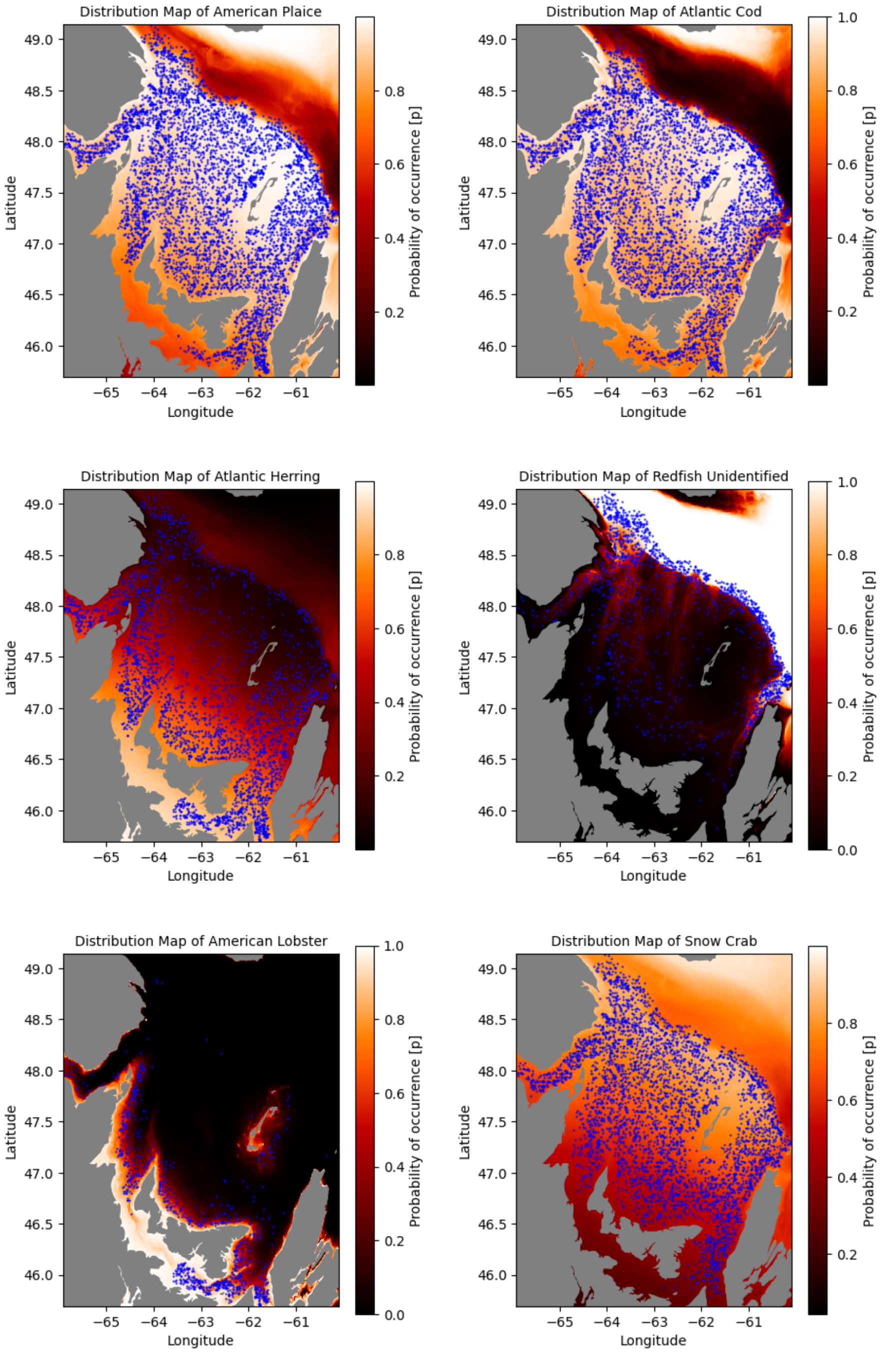


Figure 8: The receiver operator curves (ROC) for each species model. The area under the curve (AUC), Youden's J threshold and max MCC threshold are shown for each graph. These graphs display the trade-off between obtaining true positives and false positives when selecting different discrimination thresholds.



	Default	Max(Spec. + Sens.)	Spec. = Sens.	Max. ACC	Max. MCC
American plaice	0.5	0.15	0.125	0.275	0.250
Atlantic cod	0.5	0.20	0.150	0.425	0.425
Atlantic herring	0.5	0.50	0.575	0.425	0.450
Redfish unidentified	0.5	0.70	0.825	0.500	0.525
American lobster	0.5	0.80	0.775	0.475	0.475
Snow crab	0.5	0.45	0.325	0.475	0.475

Figure 9: A table key discrimination thresholds for each species model.



Distribution Map of American Plaice

Distribution Map of Atlantic Cod

Distribution Map of Atlantic Herring

Distribution Map of Redfish Unidentified

Distribution Map of American Lobster

Distribution Map of Snow Crab

Figure 10: Distribution maps for all six species. Each map displays the inferred estimated probability for each point on the map and is depicted using a heatmap themed color mapping. The true occurrences for each species are overlain in order to provide a visual sense of model's performance.

4 Conclusions

4.1 Was the model useful?

Within the geographic scope of this dataset, certain species were found to be ubiquitous while others. For example, American plaice were detected at the majority of sites, while American lobster only turned up in certain geographic areas. It was interesting to note that latitude, longitude and elevation were still found to be significant predictors for each species despite those differences.

The component of the analysis focusing on discrimination thresholds was interesting but less useful. Specifically, I was surprised to see all the key thresholds for American plaice converge at 1.0.

The reason why it was not useful, is that the goal of the analysis was not to predict presence-absence, per se, but instead to produce a distribution map of based on the resulting probabilities of detection.

4.2 What did you learn about your data set?

Aquatic ecosystems face numerous challenges, from over-fishing to climate change to the introduction of invasive species. Distribution

I learn that logistic regression

seemed to be more specialized and heterogeneous in their distributions. In the case of the former, American plaice being the best example, the models were of limited use. For species like Redfish and Lobster, which are known to respond strongly to water depth, the models were seemed to be very effective. In the heatmap for Lobster, it was interesting to note the areas which the model predicted the presence to be highest are very well-established lobster fishing areas (hence why no samples were collected there).

Would be really interesting to look at the trends over time.

References

[1] NumFOCUS. Xarray. https://docs.xarray.dev/en/stable/, 2024. Accessed on 2024-04-13.

[2] Government of Canada. Open government portal. https://open.canada.ca/, 2023. Accessed on 2023-12-11.

[3] Government of Canada. Nafo division 4t groundfish research vessel trawl survey (september survey) dataset. https://open.canada.ca/data/en/dataset/1989de32-bc5d-c696-879c-54d422438e64, 2024. Accessed on 2024-04-13.

[4] Government of Manitoba. Provinces and territories of canada, april 2022. https://open.canada.ca/data/en/dataset/85efc01b-163f-ebba-2378-c43eadfb3b3f, 2022. Accessed on 2024-04-13.

[5] The General Bathymetric Chart of the Oceans (GEBCO). Gebco gridded bathymetry data download. https://download.gebco.net/, 2024. Accessed on 2024-04-13.

[6] Wikipedia. Spatial analysis. https://en.wikipedia.org/wiki/Spatial_analysis#Spatial_auto-correlation, 2024. Accessed on 2024-04-13.