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

Statistics for Data Science

David Fishman[1]

Email ids: [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 [3].

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 [**?**].

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 [3]. The name of the dataset is NAFO Division 4T groundfish research vessel trawl survey (September Survey) dataset [**?**]. 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 [**?**].

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 [**?**]. 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!

Figure 1: This figure shows the raw data, and aggregated forms of Chicken Meat, Hogs and Wheat.

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

2.4.1 Fishing Data

2.4.2 Mapping Data

2.4.3 Elevation Data

2.5 What challenges did you face?

Figure 2: A scatter plot of Meat Chicken prices vs. Meat Chicken production.

Figure 3: Results from an Ordinary Least Squares Regression performed on Chickens raised in Canada.

all of the models converged!!

3 Analysis

3.1 What trends, correlations, and/or patterns do you see in the data?

Different trends were observed for different data. This section will outline the results of our analysis for the three crops.

3.1.1 Meat Chickens

The data frame used for the final analysis contained eight observations. When the prices and production were plotted using a scatterplot, a clear linear relationship was observed. The scatterplot and trend-line for this relationship can be observed in Figure 2. Using the statsmodels package in Python, we determined the coefficients for the linear model using the ordinary least squares (OLS) approach. The output from the regression can be viewed in Figure 3.

The null hypothesis for this test was that no relationship exists between price and production. Following the regression, there was the resulting model:

$$y = 26.4810 + 50.9459x_1$$

- Where y is the number of chickens, measured in millions of individuals
- and x_1 is the price of chicken meat, in dollars per kilogram

The F-statistic for this model was observed to be 12.47; under a normal distribution the chances of observing this value are approximately 0.01%. Setting our p-value to 0.05, we would be forced to reject the null hypothesis and conclude there is indeed a relationship between the price and production of chicken in Canada. Finally, The observed value for R^2 was 0.675 which

Figure 4: A scatter plot of hog prices vs. hog production in Canada.

Figure 5: Results from an Ordinary Least Squares Regression performed on hog production in Canada.

means that approximately 67% of the variance of the data can be accounted for by this model.

3.1.2 Hogs

The data frame used for the final analysis contained a total of 38 observations. When the prices and production were plotted using a scatterplot, the ellipsoid did not appear to have any meaningful structure. The scatterplot and trend-line are displayed in Figure 4. Using the statsmodels package in Python, we determined the coefficients for the linear model using the ordinary least squares (OLS) approach. The output from the regression can be viewed in Figure 5.

The null hypothesis for this test was that no relationship exists between price and production. Following the regression, there was the resulting model:

$$y = 20.3703 + 0.0441x_1$$

- Where y is the estimated output of farms, measured in millions of individuals
- and x_1 is the price of hog, in dollars per hundredweight

The F-statistic for this model was observed to be 0.3441. With our p-value set to 0.05, the null hypothesis that no relationship exists between the price and production of hogs in Canada would be accepted. Based on the observed R^2 was 0.009, we can state that effective none of the variance in the dataset was explained by this model.

3.1.3 Wheat

The data frame used for the final analysis contained 22 observations. When the prices and production were plotted using a scatterplot, some degree of linearity was observed. The scatterplot and trend-line for this relationship can be observed in Figure 6. Using the statsmodels package in Python,

Figure 6: A scatter plot of wheat prices vs. wheat production.

Figure 7: Results from an Ordinary Least Squares Regression performed on wheat production in Canada.

we determined the coefficients for the linear model using the ordinary least squares (OLS) approach. The output from the regression can be viewed in Figure 7.

The null hypothesis for this test was that no relationship exists between price and production. Following the regression, there was the resulting model:

$$y = 1.285e4 + 41.3883x_1$$

- Where y is the production of wheat, in thousands of metric tonnes
- and x_1 is the price of wheat, in dollars per metric tonne

The F-statistic for this model was observed to be 8.065; under a normal distribution the chances of observing this value are approximately 0.01%. With a p-value set to 0.05, we would be forced to reject the null hypothesis and conclude there is indeed a relationship between the price and production of wheat in Canada, albeit not as strong of a relationship as with chicken meat. Finally, The observed value for R^2 was 0.287 which means that approximately 29% of the variance of the data can be accounted for by this model. While this finding is significant, it is clear there are more factors involved in understanding the variations of production of wheat in Canada than solely price.

4 Conclusions

4.1 What did you learn about your data set?

The major takeaway from this analysis is that the production of different crops will respond to different factors, differently. In the case of crops such as chicken meat, price appears to be a stronger driver behind crop production; when prices of meat is high, producers tend to raise more animals. The general trend behind wheat production was similar although a more significant portion of the variation in production was unaccounted for by the current price of the commodity. On the other hand, current hog prices clearly do not have an influence on animal production on hog farms; This result was unexpected, especially considering the trend detected in the production of chickens. Swine, unlike chicken meat is not under national supply management and thus more subject to fluctuation of prices due to international markets. This fact might help account for the different responses observed in the two types of meat production. Specifically, producers, in anticipation of highly variable prices, might decide to hedge their bets by investing in their farms regardless of the current state of the market.

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

References

[1] Statistics Canada. Census of agriculture. https://www.statcan.gc.ca/en/census-agriculture, Dec 2021. Accessed on 2023-12-11.

[2] NumFOCUS. pandas. https://pandas.pydata.org/pandas-docs/stable/index.html, 2023. Accessed on 2023-12-11.

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