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.

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

A basic tentant of ecology is that different species acquire different specializations and thus inhabit different ecological niches and strategies. In aquatic ecosystems, the physical and ecological environments

What is your rationale for there being a correlation in the data that you're looking to confirm and/or exploit?

The goal of this analysis is to explore and better understand changes in the production of agricultural crops in Canada. Understanding these fluctuations can yield valuable insights into the Canadian economy and the agricultural sector.

Statistics Canada collects copious amounts of data via the Census of Agriculture [1] thus making these datasets excellent candidates for analysis in our group project.

1.3 What question(s) do you want to answer?

While the collections of elements affecting the total production of agricultural products are complex and multifaceted, this report we focus solely on a single variables: price. Specifically, we seek to answer the question of what affect does price have on the production of different agricultural products?

1.4 What hypothesis(es) do you have and what is your approach to tackle the problem?

Our hypothesis is that there exists a relationship between the price of a commodity and its production in Canada, particularly for commodities without any Supply Management interference from the Government. To tackle the problem, we sourced data from the Government's Open Data Portal, prepared the data and then created and validated a model of select products. We then applied linear regression on that dataset to search for a relationship between price and production.

2 Data Preparation

2.1 What was your data source (e.g., web scraping, corporate data, a standard machine learning data set, open data, etc.)?

The dataset used in this study is sourced from open data published by the Government of Canada, on its Open Government Portal [3], an official government website. The use of open data from a reputable government source enhances transparency and allows for reproducibility in research.

2.2 How good was the data quality?

In general, the quality of the data was very good. Aside from the minimal transformations and cleaning required, the dataset was structured in a way that facilitated our analysis. Specifically, each row of data related to a single observation and each column was an attribute of that datum. Occasionally, observations had overlapping data, for example, one row containing values for a single province and another containing data for all of Canada. When this occurred, it was important care was taken not to double-count those data.

2.3 What did you need to do to procure it?

To procure the dataset, we downloaded the csv of the dataset and we ingested it using the pandas library [2].

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

First, we segmented our analysis into three distinct categories: Chicken Meat, Hogs, and Wheat. The main libraries we used for our analysis were pandas, seaborn, numpy and statsmodels.api. Utilizing a dedicated helper function, we crafted time series plots for each category, with a focus on prices —a pivotal variable in our analysis. To ensure data completeness, we applied a boolean mask to handle null values. While exploring the data we recognized the historical prices of Newfoundland is very different from other provinces as you can see in Figure 1 so we decided to exclude Newfoundland from our analysis. We decided that it would be easiest to work with the data if all the prices were aggregated into a single response so we created a time series for

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

each category where the mean prices for each category is the data and the datetime objects as the index.

2.5 What challenges did you face?

One notable hurdle emerged with certain features, like REF_DATE, initially recorded as text, necessitating a round of feature engineering to render them immediately usable. The pricing dataset, with its varied units of measurement for a single commodity, introduced complexity, prompting us to carefully address this diversity during our analysis. The Wheat production dataset presented its own intricacies, featuring multiple entries for a given year. This intricacy made comparisons with other commodities a nuanced task. Meanwhile, the diverse data ranges of farm products posed a challenge in selecting the most pertinent ones for our analytical lens. The separation of price and production datasets added layers of complexity, urging us to integrate these disparate sources for a more comprehensive understanding. Addressing null values within the merged datasets became another crucial step in ensuring the integrity of our analysis.

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.