An Aquaculture Study: Analysing the Pre-Fattening Period of Two Fish Species

Dimitrios Megkos
School of Mathematics, Computer Science and Engineering
City, University of London

Abstract: The aim of this study is to analyze 2019 and 2020 data from the pre-fattening production of two fish species, European Bass and Sea Bream, and find which species has the better survival rate, production time, and feed conversion ratio. The results revealed that Sea Bream is better regarding survival rate and feed conversion ratio, though European Bass requires less time to produce. Further research was conducted in order to find which year had the better quality of eggs and the most sales, revealing that 2020 had less sales but greater quality of eggs compared to 2019.

I. Introduction

The benefits of seafood are many, having an important role in brain and heart health. The American Heart Association recommends eating fish at least two times per week as part of a healthy diet [1]. The increasing demand for seafood over the last fifty years has contributed greatly to the rise of aquaculture production. Fisheries and aquaculture remain important sources of food, nutrition, income and livelihoods for hundreds of millions of people around the world [2]. Aquaculture accounts for 50 percent of the fish consumed globally, according to a 2009 report by an international team of researchers [3].

The farming of aquatic animals can face many challenges during the early production stages. Data Science can help marine farms overcome these challenges by analysing the ever-growing amount of data, providing insight and helping optimise production.

II. Data and research questions

A. Data

The dataset used in this study, has been provided directly from Galaxidi Marine Farm S.A., one of the oldest and most successful aquaculture companies in Greece. It contains 2019 and 2020 data from one of the earliest production stages, the pre-fattening period, for two different fish species: European Bass and Sea Bream. During this stage, the fish that have just hatched from the eggs, are split into groups and placed into tanks where they stay until they reach a certain growth point.

Each batch of eggs is identified by a unique part number which contains important information like the species and the year the eggs were produced. There are a total of thirty-seven unique part numbers. The population of European bass is split into nineteen different part numbers, while the population of Sea bream is split into eighteen different part numbers.

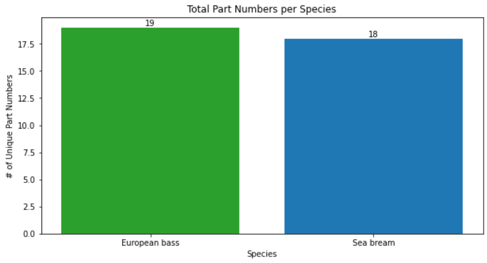

Figure 1: Bar chart showing number of unique part numbers per species.

The dataset contains 24,276 rows and 39 columns. Each row contains daily information for each tank that is being actively used for production during that period. Figure 2 shows daily tank usage for each of the two species. As each batch grows and gains weight, more tanks are being actively used. A similar pattern is identified for both species, suggesting that production is higher from December to April.

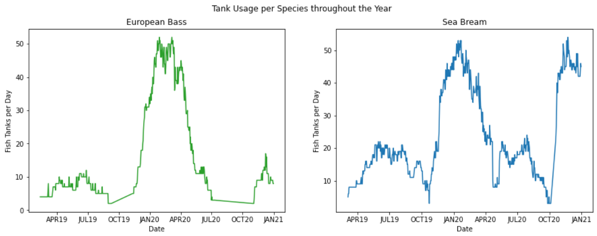

Figure 2: Line plot showing tank usage per species throughout the year.

There are several missing data that need to be dealt with and some new features will have to be created. Furthermore, for the purpose of the analysis, the entire dataset needs to be transformed so that each row contains total daily information for each unique part number, for both species.

B. Research Questions

The complexity of the pre-fattening production process for aquatic animals, has the potential to give birth to interesting questions. Since the marine farm produces two different fish species, it would be interesting to analyse each species' features and compare the differences they may have. The company has been producing and selling both species for years with great success but what if there are different challenges for each species during production time?

There are three metrics that are critical for every aquaculture company and can heavily influence decision making. The first one is the survival rate, which indicates how resilient is the fish against difficult conditions. The second one is the production time, which is the number of days a species requires to reach a certain weight point. The third one is the feed conversion ratio, which describes the amount of feed needed for the fish to gain a pound or kilogram of body weight.

Having daily data from both 2019 and 2020 gives the opportunity for some extra yearly comparisons. A holistic view of the entire year is crucial for the strategic planning of the next year. Furthermore, considering 2020 was the year the Covid-19 pandemic happened, it would be interesting to see the effects it had on the production.

This study aims to answer the following research questions:

1) Which of the two fish species is better regarding a) survival rate b) production time c) feed conversion ratio (FCR)?

2) Which year had a) the better quality of eggs b) the most sales?

III. Analysis

In order to explore this study's research questions, an analytical plan was created, and the following steps were taken.

A. *Data preparation*

The first step of the analysis is to prepare the data and look for any missing values or values that do not provide any information.

The dataset originally contained three numeric columns that had only zero values. Because in this case the zero is meaningful, they were removed, along with eighteen more columns, due to being irrelevant to the analysis. Furthermore, there were five columns that contained null values which were all removed except for one, random sampling_gr.

Random sampling_gr represents the weight in grams of a sample fish chosen randomly from a tank on a certain day. It plays an important role to one part of the analysis, since it will be used to calculate the total batch biomass for each day. A null value means that there was no sampling for that day, therefore all null values were replaced with zero number. In addition, there were several days where not all tanks were included in the sampling process of some specific batches therefore, the true total batch biomass for those days could not be calculated. For that reason, random sampling_gr on those days was replaced by zero.

B. *Data derivation*

The next step of the analysis is to transform the dataset and create new information from pre-existing data.

Using the random sampling_gr data, where available, a new column was created containing the daily total biomass of each fish tank, by multiplying random sampling_gr with fish number and dividing by one thousand to convert grams to kilograms.

Initially, each row in the dataset contained daily information for each tank. For the purpose of the analysis, it was transformed so that each row contains total daily information for each unique part number, for both species.

By utilizing the transformed dataset and the newly created biomass feature, a new column was created containing the daily mean stock weight in grams for each unique part number, by dividing biomass by the number of fish. This information is required to answer parts two and three of the first research question. Since the daily mean stock weight of each batch of fish is available only for a limited number of days, a simple linear regression model was created to estimate the numbers for the rest of the days.

In order to answer parts one and two of the first question, four new data frames were created, two for each species, so they could be analysed individually and compared with each other. Each row of these data frames contains consolidated data for each unique part number. A fifth data frame was created containing the average feed conversion ratio for both species, required to investigate part three of the first question.

A sixth and final data frame was created to answer both parts of the second question. Each part number contains the year the eggs were produced. By extracting this information, each batch of fish was grouped together based on the year they were produced in order to do yearly comparisons.

C. *Construction of models*

The construction of a model proved to be crucial for this study. Estimating the missing mean stock weight values for every unique part number was mandatory for parts two and three of the first question. All the steps below were repeated twice, once for each species.

The first step was to find the correlation of mean stock weight with the rest of the features. Figure 3 shows a high correlation with Fish age for both species. As the fish are getting older, they gain more weight.

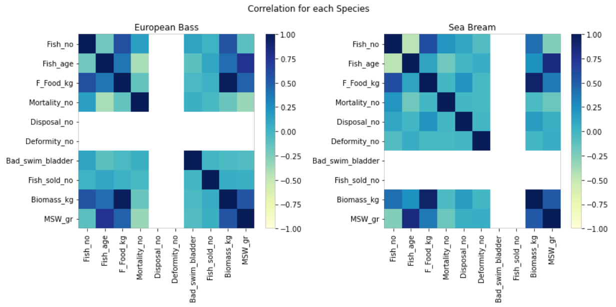

Figure 3: Pearson correlation heatmap for both species.

The next step was to identify the function that best describes mean stock weight and Fish age. Figure 4 shows that the target variable follows an exponential function, having a right skew.

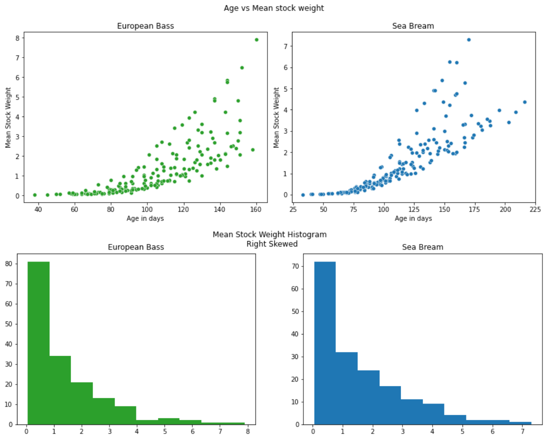

Figure 4: Scatterplot and histogram showing the distribution of mean stock weight for both species.

During this step, a few outliers were revealed. Using the boxplots shown in Figure 5 and by applying logarithm to mean stock weight, they were removed.

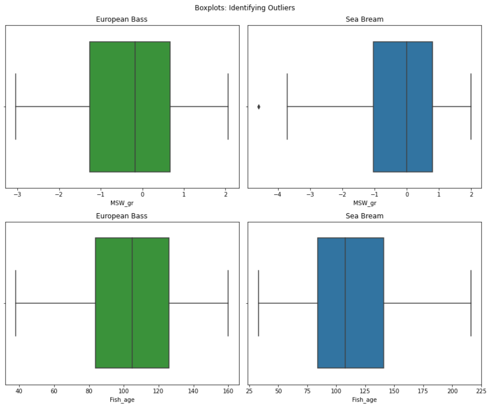

Figure 5: Boxplots for mean stock weight and fish age for both species.

Using Q-Q plot as an extra verification step, it was confirmed that mean stock weight follows an exponential function, as shown in Figure 6 below.

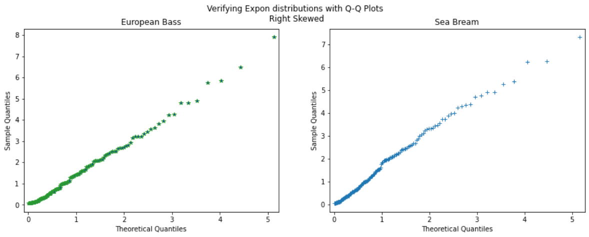

Figure 6: Q-Q plot verifying exponential function for both species.

By applying logarithm to mean stock weight, the exponential function was transformed into a linear function with one variable. This transformation simplified the modeling process, since linear regression models are easy to implement.

Since each unique batch of fish is isolated from each other, there are minor differences in growth rate, even within the same species. For that reason, one linear regression model was created for each unique part number. However, this approach also greatly reduced the amount of available training data for each model. This problem was overcome by implementing K-Fold cross-validation [4] during the training process, which greatly improved regression metrics.

Lastly, exponentiation was applied to each model's output value, in order to get the final mean stock weight estimations. Figure 7 shows the residuals of two models, one for each species.

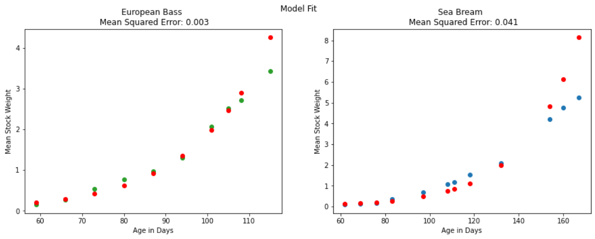

Figure 7: Linear regression residuals with mean square errors for both species.

D. *Validation of results*

Originally, one linear regression model was used for each species in order to estimate the mean stock weight. However, this method proved to be wrong, since each batch of fish have different growth rate. As a result, for some part numbers the mean stock weight was decreasing as the fish were getting older. This was discovered by exploring the data frame after including the estimations provided by the models. For this reason, it was decided that a different linear regression model would be used for each unique part number.

IV. Findings, Reflections and Further Work

The analytical plan that was created and the steps that were taken proved to be successful in answering this study's research questions.

A. *Findings*

Which of the two fish species has the better survival rate?

The sum of all the fish that died during the production was compared with the sum of the total number of fish, for both species. Figure 8 shows that Sea Bream is better regarding how much stock it is left at the end of the pre-fattening production stage. More specifically, only 6.1% of the total Sea Bream population died during production, compared to 14.6% of European Bass population.

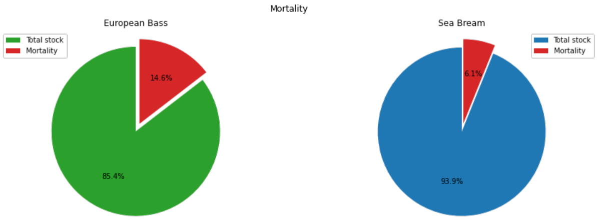

Figure 8: Pie chart showing the mortality percentage for both species.

Which of the two fish species has the better production time?

In order to find which species has the better production time, a common Mean stock weight target must be set. The target is going to be around 2.1 grams, since it represents more than 80% of the data, according to Figure 9 below.

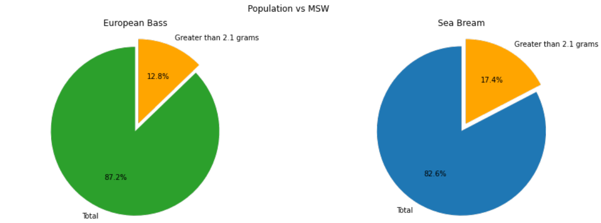

Figure 9: Pie chart showing percentage of fish that are greater than 2.1 grams.

For each part number, the number of days taken to reach around 2.1 gram of weight was calculated, and the average was used to find the production time. Figure 10 shows that it takes on average fourteen less days for European Bass to reach approximately two grams of weight compared to Sea Bream, therefore, European Bass has better production time than Sea Bream.

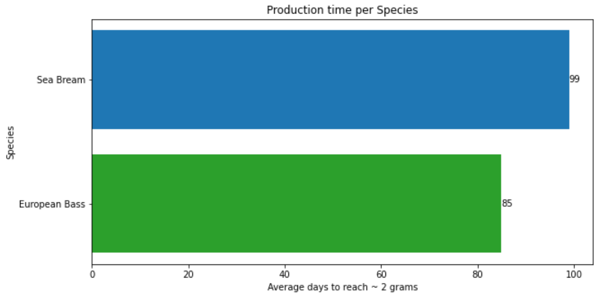

Figure 10: Bar chart showing production time for both species.

Which of the two fish species has the better feed conversion ratio?

The Feed conversion ratio [5] is the total feed consumed divided by the total weight of product produced, which is the final weight of the product minus the starting weight of the product. The two fish species have very similar Feed Conversion Ratio numbers, more specifically European Bass's Feed Conversion Ratio is 13.84, while Sea Bream's Feed Conversion Ratio is 14.47. Although a clear winner cannot be declared, European Bass has a smaller, therefore better Feed Conversion Ratio than Sea Bream.

Which year had the better quality of eggs?

The year that had the lowest percentage of mortality and bad swim bladder was considered the year with the better quality of eggs. Figure 11 shows that the year 2019 had 1.4% less fish suffering from bad swim bladder, though the mortality was 3.8% higher than the year 2020. Although the differences are not significant, the year 2020 had the better figures, therefore the better quality of eggs.

Which year had the most sales?

Figure 11 shows that the year 2020 had significantly less sales compared to the year 2019. There is a high possibility that the reason behind the decrease in sales was the Covid-19 pandemic, though there are not enough data to back this claim and find the exact cause of this phenomenon.

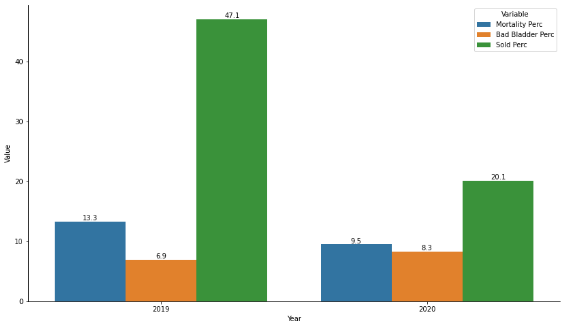

Figure 11: Bar chart showing yearly percentages of mortality, bad swim bladder and sales.

B. *Reflections*

Regarding this study's first question, though Sea Bream had the least losses and the better feed conversion ratio, European Bass requires less time, therefore less resources to produce. Both species have positives and negatives, explaining why the company keeps producing both.

Year 2020 was a good year regarding the quality of the eggs. Although the Covid-19 pandemic was possibly the main suspect for the sales drop, it had no effect on the pre-fattening production.

C. *Further work*

This study analysed the pre-fattening production stage and revealed some intriguing patterns. It would be interesting, as a further work, to collect more data and explore the next production stage in order to do more complex comparisons between production stages.

Word counts

- Abstract: 97 words
- Introduction: 134 words
- Data: 259 words
- Research questions: 254 words
- Analysis: 917 words
- Findings, reflections and further work: 565 words

References

[1] The American Heart Association. “Fish and Omega-3 Fatty Acids” Heart.org. Heart.org, 1 November 2021.

[2] FAO. 2016. The State of World Fisheries and Aquaculture 2016. Contributing to food security and nutrition for all. Rome. 200 pp

[3] Stanford University. "Half Of Fish Consumed Globally Is Now Raised On Farms, Study Finds." ScienceDaily. ScienceDaily, 8 September 2009.

[4] Jason Brownlee. (2020). Repeated k-Fold Cross-Validation for Model Evaluation in Python. Machine Learning Mastery.

[5] Stickney, Robert R. (2009) Aquaculture: An Introductory Text, page 248, CABI, ISBN 9781845935894.